Analysis of optimal differential gene expression

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biophysik/Theoretische Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herr Dipl.-Phys. Wolfram Liebermeister
geboren am 25.7.1972 in Tübingen

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jürgen Mlynek

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Michael Linscheid

Gutachter:

1. Prof. Dr. Reinhart Heinrich
2. Prof. Dr. Thomas Höfer
3. Prof. Dr. Martin Vingron

eingereicht am:	21. Mai 2003
Tag der mündlichen Prüfung:	13. Januar 2004

Abstract

This thesis is concerned with the observation that coregulation patterns in gene expression data often reflect functional structures of the cell. First, simulated gene expression data and expression data from yeast experiments are studied with independent component analysis (ICA) and with related factor models. Then, in a more theoretical approach, relations between gene expression patterns and the biological function of the genes are derived from an optimality principle.

Linear factor models such as ICA decompose gene expression matrices into statistical components. The coefficients with respect to the components can be interpreted as profiles of hidden variables (called "expression modes") that assume different values in the different samples. In contrast to clusterings, such factor models account for a superposition of effects and for individual responses of the different genes: each gene profile consists of a superposition of the expression modes, which thereby account for the common variation of many genes. The components are estimated blindly from the data, that is, without further biological knowledge, and most of the methods considered here can reconstruct almost sparse components. Thresholding a component reveals genes that respond strongly to the corresponding mode, in comparison to genes showing differential expression among individual samples.

In this work, different factor models are applied to simulated and experimental expression data. To simulate expression data, it is assumed that gene expression depends on several unobserved variables ("biological expression modes") which characterise the cell state and that the genes respond to them according to nonlinear functions called "gene programs". Is there a chance to reconstruct such expression modes with a blind data analysis? The tests in this work show that the modes can be found with ICA even if the data are noisy or weakly nonlinear, or if the numbers of true and estimated components do not match. Regression models are fitted to the profiles of single genes to explain their expression by expression modes from factor models or by the expression of single transcription factors. Nonlinear gene programs are estimated by nonlinear ICA: such effective gene programs may be used for describing gene expression in large cell models. ICA and similar methods are applied to expression data from cell-cycle experiments: besides biologically interpretable modes, experimental artefacts, probably caused by hybridisation effects and contamination of the samples, are identified. It is shown for single components that the coregulated genes share biological functions and the corresponding enzymes are concentrated in particular regions of the metabolic network.

Thus the expression machinery seems to portray - as an outcome of evolution - functional relationships between the genes: regarding the economy of resources, it would probably be inefficient if cooperating genes were not coregulated. To formalise this teleological view on gene expression, a mathematical model for the analysis of optimal differential expression (ANODE) is proposed in this work: the model describes regulators, such as genes or enzymes, and output variables, such as metabolic fluxes. The system´s behaviour is evaluated by a fitness function, which, for instance, rates some of the metabolic fluxes in the cell and which is supposed to be optimised. This optimality principle defines an optimal response of regulators to small external perturbations. For calculating the optimal regulation patterns, the system to be controlled needs to be known only partially: it suffices to predefine its possible behaviour around the optimal state and the local shape of the fitness function. The method is extended to time-dependent perturbations: to describe the response of metabolic systems to small oscillatory perturbations, frequency-dependent control coefficients are defined and characterised by summation and connectivity theorems. For testing the predicted relation between expression and function, control coefficients are simulated for a large-scale metabolic network and their statistical properties are studied: the structure of the control coefficients matrix portrays the network topology, that is, chemical reactions tend to have little control on distant parts of the network. Furthermore, control coefficients within subnetworks depend only weakly on the modelling of the surrounding network.

Several plausible assumptions about appropriate expression patterns can be formally derived from the optimality principle: the main result is a general relation between the behaviour of regulators and their biological functions, which implies, for example, the coregulation of enzymes acting in complexes or functional modules. In this context, the functions of genes are quantified by their linear influences (called "response coefficients") on fitness-relevant cell variables. For enzymes controlling metabolism, the theorems of metabolic control theory lead to sum rules that relate the expression patterns to the structure of the metabolic network. Further predictions concern a symmetric compensation for gene deletions and a relation between gene expression and the fitness loss caused by gene deletions. If optimal regulation is realised by feedback signals between the cell variables and the regulators, then functional relations are also portrayed in the linear feedback coefficients, so genes of similar function may be expected to share inputs from the same signalling cascades. According to the model of optimal regulation, expression profiles are linear combinations of response coefficient profiles: tests with experimental expression profiles and simulated control coefficients support this hypothesis, and the common components which are estimated from both kinds of data provide a vivid picture of the metabolic adaptations that are required in different environments.

To summarise, empirical relations between gene expression and function have been confirmed in this work. Furthermore, such relations have been predicted on theoretical

grounds. A main aim is to clarify teleological assertions about gene expression by deriving them from explicit assumptions, and thus to provide a theoretical framework for the integration of expression data and functional annotations. While other authors have compared expression to functional gene categories or topologically defined metabolic pathways, I propose to relate it to the response coefficients. A main result of this work is that general relations are predicted between a gene's function, its optimal expression behaviour, and its regulatory program. Where the assumption of optimality is valid, the model justifies the use of expression data for functional annotation and pathway reconstruction, and it provides a function-related interpretation for the linear components behind expression data. The methods from this work are not limited to gene expression data: the factor models are applicable to protein and metabolite data as well, and the optimality principle may also apply to other regulatory mechanisms, such as the allosteric control of enzymes.

Keywords:
differential expression, optimal control, metabolic control theory, gene function

Zusammenfassung

Diese Doktorarbeit behandelt die Beobachtung, daß Koregulationsmuster in Genexpressionsdaten häufig Funktionsstrukturen der Zelle widerspiegeln. Zunächst werden simulierte Genexpressionsdaten und Expressionsdaten aus Hefeexperimenten mit Hilfe von Independent Component Analysis (ICA) und verwandten Faktormodellen untersucht. In einem eher theoretischen Zugang werden anschließend Beziehungen zwischen den Expressionsmustern und der biologischen Funktion der Gene aus einem Optimalitätsprinzip hergeleitet.

Lineare Faktormodelle, beispielsweise ICA, zerlegen Genexpressionsmatrizen in statistische Komponenten: die Koeffizienten bezüglich der Komponenten können als Profile von verborgenen Variablen ("Expressionsmoden") interpretiert werden, deren Werte zwischen den Proben variieren. Im Gegensatz zu Clustermethoden beschreiben solche Faktormodelle eine Überlagerung biologischer Effekte und die individuellen Reaktionen der einzelnen Gene: jedes Genprofil besteht aus einer Überlagerung der Expressionsmoden, die so die gemeinsamen Schwankungen vieler Gene erklären. Die linearen Komponenten werden blind, also ohne zusätzliches biologisches Wissen, aus den Daten geschätzt, und die meisten der hier betrachteten Methoden erlauben es, nahezu schwach besetzte Komponenten zu rekonstruieren. Beim Ausdünnen einer Komponente werden Gene sichtbar, die stark auf die entsprechende Mode reagieren, ganz in Analogie zu Genen, die differentielle Expression zwischen einzelnen Proben zeigen.

Verschiedene Faktormodelle werden in dieser Arbeit auf simulierte und experimentelle Expressionsdaten angewendet. Bei der Simulation von Expressionsdaten wird angenommen, daß die Genexpression von einigen unbeobachteten Variablen ("biologischen Expressionsmoden") abhängt, die den Zellzustand beschreiben und deren Einfluss auf die Gene sich durch nichtlineare Funktionen, die sogenannten Genprogramme, beschreiben läßt. Besteht Hoffnung, solche Expressionsmoden durch blinde Datenanalyse wiederzufinden? Die Tests in dieser Arbeit zeigen, daß die Moden mit ICA recht zuverlässig gefunden werden, selbst wenn die Daten verrauscht oder leicht nichtlinear sind und die Anzahl der wahren und der geschätzten Komponenten nicht übereinstimmt. Regressionsmodelle werden an Profile einzelner Gene angepasst, um ihre Expression durch Expressionsmoden aus Faktormodellen oder durch die Expression einzelner Transkriptionsfaktoren zu erklären. Nichtlineare Genprogramme werden mit Hilfe von nichtlinearer ICA ermittelt: solche effektiven Genprogramme könnten zur Beschreibung von Genexpression in großen Zellmodellen Verwendung finden. ICA und verwandte Methoden werden auf Expressionsdaten aus Zellzyklusexperimenten angewendet: neben biologisch interpretierbaren Moden

werden experimentelle Artefakte identifiziert, die vermutlich Hybridisierungseffekte oder eine Verunreinigung der Proben widerspiegeln. Für einzelne Komponenten wird gezeigt, daß die koregulierten Gene gemeinsame biologische Funktionen besitzen und daß die entsprechenden Enzyme bevorzugt in bestimmten Bereichen des Stoffwechselnetzes zu finden sind.

Die Expressionmechanismen scheinen also - als Ergebnis der Evolution - Funktionsbeziehungen zwischen den Genen widerzuspiegeln: es wäre unter ökonomischen Gesichtspunkten vermutlich ineffizient, wenn kooperierende Gene nicht auch koreguliert würden. Um diese teleologische Vorstellung von Genexpression zu formalisieren, wird in dieser Arbeit ein mathematisches Modell zur Analyse der optimalen differentiellen Expression (ANODE) vorgeschlagen: das Modell beschreibt Regulatoren, also beispielsweise Gene oder Enzyme, und die von ihnen gesteuerten Variablen, zum Beispiel metabolische Flüsse. Das Systemverhalten wird durch eine Fitnessfunktion bewertet, die beispielsweise von bestimmten Stoffwechselflüssen abhängt und die es zu optimieren gilt. Dieses Optimalitätsprinzip definiert eine optimale Reaktion der Regulatoren auf kleine äußeren Störungen. Zur Berechnung optimaler Regulationsmuster braucht das zu regulierende System nur teilweise bekannt zu sein: es genügt, sein mögliches Verhalten in der Nähe des optimalen Zustandes sowie die lokale Form der Fitnesslandschaft zu kennen. Die Methode wird auf zeitabhängige Störungen erweitert: um die Antwort von Stoffwechselsystemen auf kleine oszillatorische Störungen zu beschreiben, werden frequenzabhängige Kontrollkoeffizienten definiert und durch Summations- und Konnektivitätstheoreme charakterisiert. Um die vorhergesagte Beziehung zwischen Expression und Funktion zu prüfen, werden Kontrollkoeffizienten für ein großes Stoffwechselnetz simuliert, und ihre statistischen Eigenschaften werden untersucht: die Struktur der Kontrollkoeffizientenmatrix bildet die Netztopologie ab, das bedeutet, chemische Reaktionen haben gewöhnlich einen geringen Einfluss auf weit entfernte Teile des Netzes. Außerdem hängen die Kontrollkoeffizienten innerhalb eines Teilnetzes nur schwach von der Modellierung des umgebenden Netzes ab.

Verschiedene plausible Annahmen über sinnvolle Expressionsmuster lassen sich formal aus dem Optimalitätsprinzip herleiten: das Hauptergebnis ist eine allgemeine Beziehung zwischen dem Verhalten und der biologischen Funktion von Regulatoren, aus der sich zum Beispiel die Koregulation von Enzymen in Komplexen oder Funktionsmodulen ergibt. Die Funktionen der Gene werden in diesem Zusammenhang durch ihre linearen Einflüsse (die sogenannten Responsekoeffizienten) auf fitnessrelevante Zellvariable beschrieben. Für Stoffwechselenzyme werden aus den Theoremen der metabolischen Kontrolltheorie Summenregeln hergeleitet, die die Expressionsmuster mit der Struktur des Stoffwechselnetzes verknüpfen. Weitere Vorhersagen betreffen eine symmetrische Kompensation von Gendeletionen und eine Beziehung zwischen Genexpression und dem Fitnessverlust aufgrund von Deletionen. Wenn die optimale Steuerung durch eine Rückkopplung zwischen Zellvariablen und den Regulatoren verwirklicht ist, dann spiegeln sich funktionale Beziehungen auch in

den Rückkopplungskoeffizienten wider. Daher liegt es nahe, daß Gene mit ähnlicher Funktion durch Eingangssignale aus denselben Signalwegen gesteuert werden. Das Modell der optimalen Steuerung sagt voraus, daß Expressionsprofile aus Linearkombinationen von Responsekoeffizientenprofilen bestehen: Tests mit experimentellen Expressionsdaten und simulierten Kontrollkoeffizienten stützen diese Hypothese, und die gemeinsamen Komponenten, die aus diesen beiden Arten von Daten geschätzt werden, liefern ein anschauliches Bild der Stochwechselvorgänge, die zur Anpassung an unterschiedliche Umgebungen notwendig sind.

Alles in allem werden in dieser Arbeit empirische Beziehungen zwischen der Expression and der Funktion von Genen bestätigt. Darüber hinaus werden solche Beziehungen aus theorischen Gründen vorhergesagt. Ein Hauptziel ist es, teleologische Aussagen über Genexpression auf explizite Annahmen zurückzuführen und dadurch klarer zu formulieren, und so einen theoretischen Rahmen für die Integration von Expressionsdaten und Funktionsannotationen zu liefern. Während andere Autoren die Expression mit Funktionskategorien der Gene oder topologisch definierten Stoffwechselwegen verglichen haben, schlage ich vor, die Funktionen von Genen durch ihre Responsekoeffizienten auszudrücken. Als ein Hauptergebnis dieser Arbeit werden allgemeine Beziehungen zwischen der Funktion, der optimalen Expression und dem Programm eines Gens vorhergesagt. Soweit die Optimalitätsannahme gilt, rechtfertigt das Modell die Verwendung von Expressionsdaten zur Funktionsannotation und zur Rekonstruktion von Stoffwechselwegen und liefert außerdem eine funktionsbezogene Interpretation für die linearen Komponenten in Expressionsdaten. Die Methoden aus dieser Arbeit sind nicht auf Genexpressionsdaten beschränkt: die Faktormodelle lassen sich auch auf Protein- und Metabolitdaten anwenden, und das Optimalitätsprinzip könnte ebenfalls auf andere Steuerungsmechanismen angewendet werden, beispielsweise auf die allosterische Steuerung von Enzymen.

Schlagwörter:
Differentielle Expression, Optimale Steuerung, Metabolische Kontrolltheorie, Genfunktion

Danksagung

Ich möchte zuallererst meinen Betreuern Reinhart Heinrich, Edda Klipp und Hans Lehrach danken, nicht nur für ihre Unterstützung bei dieser Arbeit, sondern auch für das schöne und interessante Umfeld, in dem ich die letzten Jahre über lernen und arbeiten konnte. Die Zeit in den Arbeitsgruppen an der Humboldt-Universität und im Max-Planck-Institut für molekulare Genetik war für mich eine wirkliche Bereicherung, und der Abschied von meinen Mitarbeitern fällt mir ausgesprochen schwer. Auch zahlreiche auswärtige Leute haben mir durch Gespräche bei meinem Einstieg in die Welt der Biologie geholfen und mir immer wieder neue Anregungen gegeben. Besonders erwähnen möchte ich hier Frank Bruggeman, Hanspeter Herzel, Femke Mensonides, Ralf Mrowka, Axel Nagel, Steffen Schulze-Kremer, Stefan Schuster und Rainer Spang. Ein besonderer Dank gilt auch den Kollegen, deren Daten und Algorithmen ich für meine Arbeit verwenden konnte, insbesondere A. Hyvärinen.

Contents

Chapter 1

Introduction

1.1 About this work

1.1.1 Motivation

Together with the physical structure of cells, their functional structure has been shaped by evolution, and the same can be assumed for regulatory processes which are responsible for coordinating the cell's various actions [66]. This thesis is mainly concerned with the control of gene expression, which is involved in many cell processes and responses to external stimuli (see Figure 1.1). It is studied how and why expression profiles tend to portray functional structures of the cell. Genome-wide expression can be monitored by measurements with microarrays, to study which genes respond to certain experimental interventions, which of the genes show characteristic differences between cell types, and which groups of genes show a concerted effort in their expression behaviour. On the one hand, this knowledge may help to find out details about the physical mechanisms of gene expression: which transcription factors, which other signals influence a gene, and how does it respond to them? On the other hand, the cell's expression machinery can be used as a measuring device to study the functional structure of the cell, as genes showing similar expression patterns also have a tendency to share biological functions [111] [7]. The objective of this thesis is to study these functional aspects of gene expression in more detail. I shall follow two tracks: first, coregulation structures are extracted from gene expression data by factor models. Secondly, it is assumed that the cell aims to maximise a biological fitness function and that the differential expression patterns are shaped to achieve this goal. It turns out that the resulting optimal expression patterns reflect the functions of genes, as it has been found in reality.

In many publications on gene expression, the data are organised by clusterings [26] [5], where genes and experiments with similar expression profiles are arranged into groups.

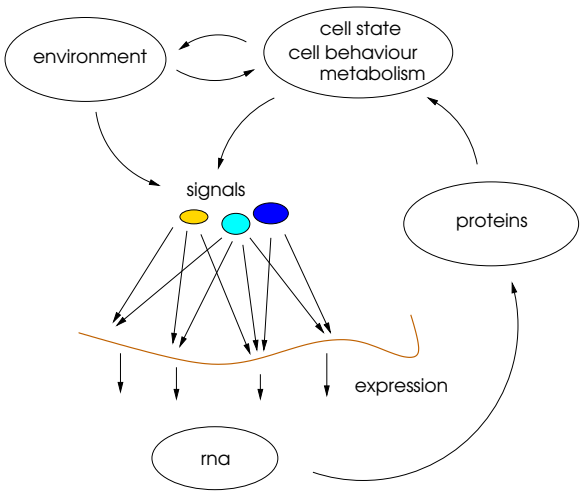


Figure 1.1: Gene expression as a part of the cell's regulatory network. Gene expression involves the production of messenger RNA, which is translated into proteins. Via the proteins, genes control the structure and the behaviour of the cell. Stimuli from the environment and processes within the cell send signals to the gene expression machinery. The resulting feedback loops can ensure homoeostasis, but may also respond specifically to external stimuli, for instance, to stress conditions.

Biclustering [110] allows for a partial coregulation, detecting genes with similar profiles in subgroups of the samples. In contrast, I shall analyse expression data by linear factor models such as independent component analysis (ICA, [84] [49] [78]). Linear factor models represent the correlation structure of multivariate data by decomposing them into linear components with predefined statistical properties. A similar decomposition is used in physics, for instance, when the dynamics of a string is described by the simple behaviour of harmonics, and in metabolic control theory, where metabolic flux distributions can be superposed from elementary modes describing metabolic pathways. Compared to clusterings, linear factor models have the advantage that each gene's expression is explained by an individual superposition of different effects, similar to the process of transcription, where a gene may be controlled by more than just one transcription factor. The corresponding linear transformation provides a set of general basic profiles from which the gene profiles can be superposed. In contrast to clusterings, where the expression profiles of genes are compared as a whole, genes will be regarded as coregulated here if they both respond strongly to the same expression mode, even if, due to other expression modes, the gene profile differ from each other.

Biologically, the individual yet concerted expression of genes can be attributed to an interplay of regulatory mechanisms: in a schematical picture, the cell state can be represented by variables called "expression modes", and each gene is controlled by a combinatorial

function (which I shall call the "gene program"), describing how the gene's expression depends on these modes. Bussemaker et al. [9] explained expression data by a linear model where each factor influences genes that share a particular sequence motif in their regulatory regions. However, the abovementioned modes need not represent biological regulators, such as transcription factors, but may also describe the cell state in a global manner. In its mathematical form, this schematic model of expression resembles the linear models for data analysis mentioned above. The statistical factor models considered in this work estimate their components blindly from the structures in the expression data, that is, no additional biological information is used. The linear model should determine a data subspace containing the systematic effects, that is, biological processes and possibly experimental artefacts, while the remaining subspace is supposed to describe weak statistical noise. A standard method for this dimension reduction is principal component analysis [22]. Furthermore, the relevant subspace should be decomposed into the separate causes of variation. This task is much more subtle: whether a method can separate distinct biological effects depends on the statistical assumptions made about the components. In this work, linear methods based on different statistical assumptions will be tested, most of which are sensitive to sparse components. As data sets, I chose publicly available data from yeast experiments, where expression during the cell cycle [107], after environmental changes [18] [10] [32], and after gene deletions [51] [61] had been studied (see Table B.3 in the appendix). For testing purposes, also simulated expression data are analysed.

As mentioned above, this work focuses on the functional aspects of gene expression: gene function can be defined qualitatively by assigning genes to functional classes, for instance to MIPS functional categories [82], gene ontology annotations [14], or KEGG metabolic pathways [64]. Such annotations state in which cellular processes or subsystems a gene is involved. Evidence for shared function can also come from associations between the gene products, for instance, for proteins forming complexes [33], interaction [112] or fusion pairs [115], or protein pairs defined by phylogenetic correlation [91] or anticorrelation. One reason to cluster gene expression data is to determine functionally related genes: in fact, expression clusters are enriched in genes from particular functional classes [111], and genes in metabolic pathways were found to be coregulated [18]. Accordingly, expression data have been used to discriminate genes with respect to functional annotations [7] [110], to predict functions [80], or to reconstruct probable metabolic pathways [118] [38] [71].

I already stated that linear models can identify components behind gene expression data which describe concerted actions of many genes. This correlated expression might be attributed to common regulation, but the components also seem to represent common biological gene functions. A particular gene expression pattern, for instance, may be attributed to a *causa efficiens*, such as a signalling pathway, which physically influences the transcript levels. Accordingly, expression data have been used to identify regulatory motifs [6] [9]. On the other hand, expression may be explained by a *causa finalis*, namely the fact

that the gene products are needed by the cell under the given conditions. Biologists often tacitly presume a form of teleonaturalism (see [3]) where "needed by the cell" translates to "increasing the cell's biological fitness, and thus selected for during evolution". In biology, optimality assumption are usually justified by evolutionary arguments: caused by mutation and selection, species change their properties and thereby increase their biological fitness, that is, the long-term reproduction rate. Under a strong selection pressure a trait is likely to achieve a local optimum. This view is the basis for modelling studies on evolutionary optimisation (see, for instance, [41] [101]). Optimality of flux distributions [24] [105] has been studied, and theoretical predictions based on optimisation could be validated by experiment [60].

The optimality-based view can be extended to regulatory systems: physically, most parts of the cell can only react to processes in their immediate neighbourhood, but their function would require a reaction to a distant process or condition. This discrepancy creates an evolutionary pressure on the development of signal-processing systems which distribute valuable information within the cell. If the cell is viewed as a machine, the signalling systems can be seen as a hard-coded implementation of a program to control it. Accordingly, the input signals of a gene should provide primarily the information about the cell state that is relevant for the gene's expression. A relation between the optimal regulation of enzymes and their control on fluxes has been derived in [67]. Optimal control of time-dependent processes [93] has been studied intensely and also been applied to the control of metabolic systems [68]. In this thesis, I shall formalise assertions about "sensible" gene expression by translating them into a mathematical framework.

It is assumed that gene regulation has to contribute to the optimisation of a given objective function. To define an optimal behaviour of genes, I shall consider how parts of the cell, such as metabolic subsystems, are supposed to behave, and how the genes contribute to this behaviour. In the framework of metabolic control theory [41] [42], the effects of gene expression on steady states of the metabolic system are described by the linear response coefficients, which can be calculated from the stoichiometry and the linearised kinetics of the chemical reactions. In this work, they are the key quantities to describe the function of genes, and it turns out that they also appear in the formulae for optimal expression patterns.

1.1.2 Structure of the text

The rest of this first chapter summarises concepts and methods from theoretical biology and statistical data analysis that are relevant for this thesis, namely the biological regulation and statistical analysis of gene expression, linear factor models, and mathematical models in cell biology. The first main topic of this thesis, the analysis of coregulation in

expression data, is treated in the chapters 2, 3, and 4. Assuming that gene expression depends on common unobserved variables, I shall study whether such variables can be estimated from microarray data. With nonlinear ICA, time courses of the expression modes are described by the components, and gene programs are reconstructed from simulated and experimental data. In chapter 4, the components are supposed to describe linear weights of the gene programs. Independent component analysis and other linear models are applied to whole-genome data to detect biologically interpretable modes and to separate them from experimental noise and artefacts. In the second main part of this work, properties of gene expression patterns are predicted from the assumption of optimal regulation. In chapter 5, a mathematical model is proposed for predicting optimal differential expression patterns. The following chapters 6 and chapter 7 add generalisations and details, for instance optimal response to oscillatory perturbations. The remaining chapters are concerned with relations between optimal expression patterns and the structure of metabolic networks. Control coefficients for a large metabolic network are calculated in chapter 8. Structural knowledge, for instance, about the metabolic network, can be used to predict properties of expression patterns. To test the predictions made, expression data are compared to simulated metabolic control coefficients in chapter 9. Chapter 10 summarises the results of this thesis. The appendices contain additional information: the data sets and algorithms used are listed in the Appendix B. The mathematical proofs and a list of mathematical symbols are given in Appendix A. Appendix C contains additional figures and further analyses of expression data.

Parts of section 4.3 and 4.4 have already been published [78] and parts of the chapters 5 and 9 are contained in [79], which has been submitted for publication. This work was supported by the Deutsche Forschungsgemeinschaft and the German Federal Ministry of Education and Research.

1.2 Gene expression

1.2.1 Mechanisms of gene expression

This section summarises basic mechanisms of gene expression in eukaryotes. A detailed description is given in [1]. The synthesis of proteins requires the production and processing of messenger RNA (mRNA), involving the following steps: an enzyme complex, the so-called called RNA polymerase, transcribes the nucleotide sequence coding for a protein to single-stranded mRNA molecules with the complementary sequence. These transcripts are spliced in the nucleus, that is, the large intron sequences are removed, and ribosomes in the cytosol translate the mRNA to proteins. After some time (usually in the order of minutes [46]), the transcripts are degraded. The correlation between RNA and protein

levels, even per gene, is not very strong [37] [61]: this may be caused by the time-delay in translation, but also by active control of RNA processing and protein decay.

Probably all steps of expression are actively controlled, in particular the initiation of transcription: in eukaryotes, a complex of general transcription factors forms at the TATA box, about 25 kilo-base-pairs (kbp) ahead of the start site, and enables the polymerase to start transcription. This process is controlled by regulatory proteins which can bind specifically to sequence motifs in the regulatory region around the start site. This region can be 50 kbp in size, much larger then a typical gene. Bound regulatory proteins can control the formation of the transcription complex even from a large distance. The regulatory proteins themselves are controlled by different mechanisms, including their own synthesis, ligand binding, phosphorylation, complex formation, unmasking, and control of their nuclear entry. Transcription also depends on the chromatin structure and on the methylation of cytosine in GC pairs, which both can be altered by regulatory proteins. In prokaryotes, a gene is usually controlled by few regulators which either activate or repress the gene, while in eukaryotes, genes are controlled by many regulators acting in a complicated combinatorial manner.

The rate of mRNA synthesis reflects the biological processes that control the initiation of transcription: the initiation rate, regarded as a mathematical function of these biochemical signals, can be supposed to have the following properties:

- **Nonlinearity**: The time-averaged initiation rate can be approximated by a smooth nonlinear function of the transcription factor concentration. If binding of a transcriptional activator is required to enable transcription, then the probability for the binding site to be occupied is a sigmoidal function of the activator's log concentration, and the gene's program is also sigmoidal.

- **Combinatorial function**: If several of the abovementioned factors are required, only certain combinations of them will trigger transcription. The resulting combinatorial function may be described by an (artificial) neural networks [22]: neural networks are schemes to build nonlinear functions by combining simple functions according to a graph structure. They are used in regression or discrimination problems in which high-dimensional nonlinear functions must be approximated by functions that are easy to parametrise. Neural networks can be fitted to data in a sequential manner (often called "learning"), in close analogy to physiological or evolutionary adaptation of signal-processing systems.

- **Modularity**: Combinatorial functions can be realised by modules in the regulatory DNA sequence. For instance (see [1]), the Drosophila morphogene "eve", which is active in 7 distinct stripes in the embryo, is regulated by seven modules: each of

them becomes activated by signals that are characteristic for a particular region in the developing embryo. Thus the pattern to be achieved is disposed by the structure of the regulatory function.

- **Time hierarchy**: Different input signals may act on different time scales. In contrast to the momentary adaptation by transcription factors, methylation of GC pairs can ensure that a gene is permanently deactivated. Hierarchical control mechanisms may manifest themselves in a hierarchy of cell states, with different cell types on top and the momentary behaviour on bottom.

Thus expression can be described by the following schematic picture: the regulatory proteins represent the state of the cell, while the corresponding binding sites on the DNA determine how genes behave in the different states. To contribute to the cell's fitness, each gene is supposed to show a particular expression behaviour under certain cell conditions, and probably, the combinatorial regulatory functions of genes (which will be called "gene programs" in this text) have evolved along with the required behaviour of the protein. Replication and modification of small sequence motifs (comparable to genetic programming [70]) provide an efficient way to create and adapt complex regulatory functions during evolution.

1.2.2 Methods for studying gene expression data

With many genes, the expression level differs among cell types, developmental stages, and external conditions, and this differential expression can be measured in parallel and on a genomic scale by the use of macro- or microarrays: the mRNA is extracted from a cell or tissue sample and transcribed to complementary DNA (cDNA), labelled by a fluorescent dye or by a radioactive isotope. This target cDNA is then hybridised to spots of probe cDNA or oligonucleotides on nylon filters or glass slides. The bound cDNA is quantified by scanning the fluorescence, yielding an intensity value for each spot, that is, for each gene represented in the array. The observed intensity is supposed to represent the mRNA concentration in the original sample, but the data may also be superposed by statistical noise and by artefacts originating from the sample preparation and the hybridisation procedure. A typical microarray contains several thousand spots, so all known yeast genes can be represented on one chip. Microarray data allow for studying details of the the regulatory system, but they also provide important diagnostic information, for instance for the discrimination of cancer types. Microarrays are not only used for measuring expression, but also for sequencing, to identify binding of transcription factors [76], single nucleotide polymorphisms, and to quantify genetically labelled yeast strains [35].

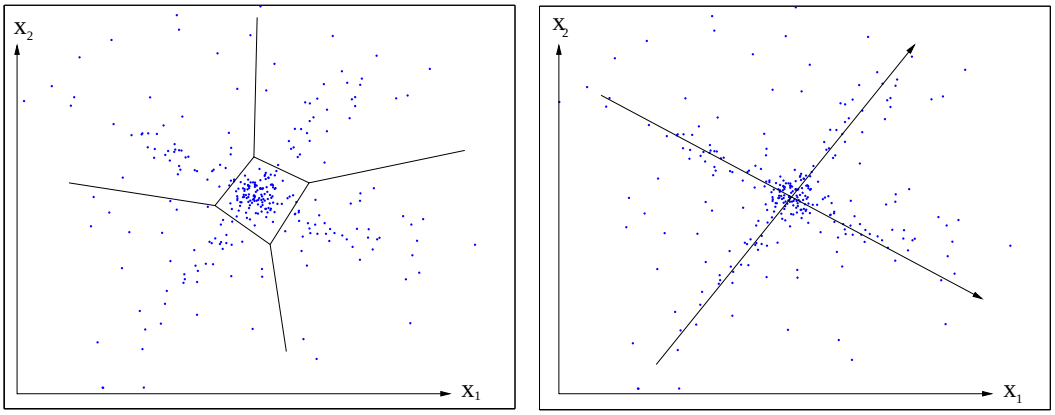


Figure 1.2: Clustering and linear factor model for multivariate data. The gene profiles are represented by points in an n-dimensional space. Its dimensions represent the expression in the different experimental samples (microarrays). In the diagrams, hypothetical data are projected to coordinates x_1 and x_2, which are linear combinations of the original samples. The data cloud consists of a central region of high density, surrounded by several "arms". Clustering and factor models describe the data cloud by structures that are adapted to the cloud's shape. Left: Clustering. The genes are grouped into different classes, according to the similarities between their profiles. Right: Linear factor model. The data space is parametrised by new coordinates. The new basis vectors have been chosen according to the shape of the data cloud.

Microarray data form a matrix, containing the gene profiles in its rows, while the experimental samples (microarrays) are represented by columns. The number of genes (typically several thousand) exceeds the number of samples (usually less than a hundred). Usually, the measured intensities vary over several orders of magnitude, and both additive and multiplicative measurement errors are present: the noise level is often as high as the typical differential expression among samples [23]. Systematic errors can be reduced by an accurate data normalisation [102], while statistical errors can be treated either by averaging over them or by estimating them as part of a statistical model. Due to the high costs, the measurements are often not repeated and usually, the noise level remains high. This may cause severe problems because genes of interest (e.g oncogenes) may show small differential expression and remain hidden in the noise.

It is common to represent expression values by their logarithms because the error distributions then become closer to normal. As the hybridisation process is not yet fully understood, it is difficult to measure absolute RNA concentrations by microarrays, and it is common to report differential expression [13] between samples or groups of samples. An important application is to determine marker genes whose expression differs between different tumour types. For each gene, a p-value denotes the probability that the degree

of differential expression observed would occur by chance. For calculating the p-value, the biological or experimental fluctuations have to be studied by comparing repeated measurements of identical samples.

Besides differential expression and coregulation of particular genes, global patterns can be studied in expression data. In Figure 1.2, hypothetical gene expression profiles, represented by points, form a data cloud which has been projected to two dimensions x_1 and x_2. If the gene profiles are centred, then for each gene, the variance over the experiments is given by the squared distance from the cloud centre, and the correlation between two genes equals the cosine of the angle between them. Multivariate data of high dimensionality require methods to extract and display information about their correlation structure. Clustering of genes determines distinct groups of genes with similar expression profiles. K-means clustering [22], for instance, is based on the assumption that the probability density behind the data is a mixture of isotropic Gaussian distributions, each giving rise to a cluster of data points. A number k of clusters is chosen, and the cluster centres and their members are estimated in parallel by maximum likelihood: practically, the average Euclidean distance between each data point and its respective cluster centre is minimised. If other distance measures than the Euclidean distance are given, each cluster can be represented by a data point (a so-called prototype vector), which is, on average, closest to the other points of its cluster. Hierarchical clustering [26] is a popular method to arrange genes and samples according into cluster trees. Linear factor models, which are explained in more detail in the following section, assume that the gene expression profiles can be explained by relatively few global variables. The data space is parametrised by new coordinates (see Figure 1.2, right), and the data can be projected to fewer dimensions. With both clustering and linear models, the results depend strongly on the data metric, which defines the similarity measure between gene profiles. The metric, in turn, depends on the normalisation scheme and on the choice and statistical weights of the experimental samples. Any preprocessing of the data effectively changes the metric, and a projection to the first principal components may yield a metric in which noise effects have less weight and the biological effects become more pronounced.

1.3 Statistical factor models

This section contains a brief overview about statistical factor models. Let us consider a pair of random variables (s, x) fulfilling

$$x = f(s, p) + \eta \tag{1.1}$$

where η denotes independent Gaussian noise representing, for instance, measurement errors. The symbols x, s, and η denote vectors. If samples of both s and the x are

given, the parameters p can be estimated by regression. In contrast to that, unsupervised methods try to infer the explanatory variables from the data: stated differently, the data variables x_l are represented, according to

$$x = f(s) + \eta \tag{1.2}$$

by transformed variables s_k called "components" or "sources". By this transformation, the data may be represented by fewer variables, without loss of important information: such a dimension reduction is useful for visualising, storing, or transmitting the data, or for simplifying a further treatment, e.g., clustering or discrimination. Contrariwise, a simple sparse coding can be achieved by increasing the number of dimensions [21].

Linear models assume f to be linear, parametrised by coefficients ("loadings") A_{kl}. A data matrix X with rows and columns representing samples and variables[1], respectively, is decomposed into

$$X_{il} = \bar{\mu}_i + \sum_k S_{ik} A_{kl} + \eta_{il} \tag{1.3}$$

or, in matrix notation,

$$X = \mu + SA + \eta \tag{1.4}$$

A is called the mixture matrix or loadings matrix. If the the data have been centred by subtracting the empirical mean μ, the linear model simply reads

$$X = SA + \eta \tag{1.5}$$

1.3.1 Model fitting and validation

If constraints are imposed on S or A or if the number of variables exceeds the number of components, a decomposition $X = S\ A + \eta$ can be estimated by maximising the likelihood $p(X|S, A)$, that is, the probability to observe the data, given S and A. If the noise is independent Gaussian with zero mean and standard deviations σ_{ik}, then the log-likelihood reads

$$\log p(X|S, A) = \log p_\eta(X - S\ A) = -\frac{1}{2} \sum_{ik} \frac{(X - S\ A)_{ik}^2}{\sigma_{ik}^2} + \text{const.} \tag{1.6}$$

If all σ_{ik} are identical, the likelihood is maximised by minimising the sum of squared errors. In Bayesian statistics (see [34]), the model parameters (elements of S and A) are

[1] In the neural processing community, the convention is different, with variables in the rows and individuals in the columns.

treated as random variables drawn from a predefined prior distribution $p(S, A)$. The log posterior probability

$$\log p(S, A|X) = \log p(X|S, A) + \log p(S, A) + \text{const.} \tag{1.7}$$

describes the probability of certain choices of S and A after observing the data. If the prior distribution is not known, it can be parametrised by so-called hyperparameters, which are then are also treated in the Bayesian framework.

The matrix product $S\ A$ in equation 1.6 is invariant to a transformation

$$\begin{aligned} S &\rightarrow S\ T \\ A &\rightarrow T^{-1}\ A \end{aligned} \tag{1.8}$$

and thus the result of the above maximum likelihood estimation is not unique. Also with independent component analysis (see below), the fitting criterion does not depend on the order and the signs of the components s_k. Uniqueness of the solution can be forced by standardising S and A: components can be sorted, for instance, by the data variance they explain.

If a model parameter is estimated, also the estimation error should be studied. Analytically, the variance of an estimator can only be calculated for relatively simple cases. The bootstrap [25] is a method to estimate the variance of estimated parameter for unknown, empirical distributions. In principle, the variation of the estimated parameters could be studied by fitting the model for many data sets sampled from the same distribution. The bootstrap does the same thing, using virtual datasets that have been created by resampling from the given data. For bootstrapping linear models, the solutions from different resampling runs must be comparable: therefore a reference model is defined, and each solution is rearranged to resemble this reference model as closely as possible.

The aim in discrimination and model fitting is not to explain the data as accurately as possible, but to fit a model to the unknown underlying distribution. The model should generalise well, that is, fit new data points from the same distribution, and thus allow for prediction. Over-fitting of the given data can be detected by cross-validation: the set of samples is split into two parts, the training set and the test set. The model parameters are fitted using the training set, and a prediction error is calculated from the model predictions for the test set. This procedure is repeated for different choices of the training and the test set to calculate an average prediction error. Cross-validation can be used to determine optimal model properties, such as the best number of hidden nodes in a neural network.

1.3.2 Principal components and independent components

We saw that the matrix decomposition (1.5) is not unique. If one of the matrices is given, as with the Fourier or wavelet transform, the other one can be found by maximum likelihood estimation. Alternatively, both matrices can be identified "blindly", that is, the new basis vectors (rows of A) are determined from the data. This requires, however, that certain statistical properties of the matrices have been specified in advance, or that a generative model with predefined statistical properties is fitted to the data. The rest of this section 1.3 gives an overview over different blind methods that have been applied to expression data or that are relevant for this work.

Principal component analysis (PCA) aims to determine components explaining maximal data variance. PCA can be seen as the parameter estimation of a multivariate normal distribution, with a density

$$p(x) \propto e^{-\frac{1}{2}(x-\mu)^T \ C^{-1} \ (x-\mu)} \tag{1.9}$$

where μ and C denote the mean and the covariance matrix, respectively. PCA centres and rotates the data, using the eigenvectors of C (estimated by the empirical covariance matrix) as the new basis vectors. The respective eigenvalues describe the variances along the principal components. Except for the centring, PCA is technically equivalent to a singular value decomposition (SVD). Both the components and the new basis vectors are orthogonal on each other, that is, linearly uncorrelated. Principal components are optimal for representing maximal data variance by few components, but the separate principal components need not have an interpretation.

Independent component analysis (ICA) [54] [57] assumes statistically independent components behind the data, with a distribution

$$p(s) = \Pi_i \ p_i(s_i) \tag{1.10}$$

Practically, the ICA model

$$X_{il} = \sum_k S_{ik} A_{kl}$$

splits the centred data matrix into a matrix product $X = SA$ (compare Figure 4.1 in chapter 2), subject to the condition that the statistical dependence between the columns of S be minimised. The dependence between random variables can be quantified by the mutual information $I = \sum_k H_k - H$, where H_k and H denote the entropy of the k^{th} variable and the total entropy, respectively (see section 1.4.3). As the total entropy H remains constant under linear transformations, the mutual information I can be minimised by minimising the marginal entropies H_k. Among the distributions with unit variance, the normal distribution yields the maximal entropy value H_N. ICA determines directions in

the data cloud where the distribution of the data is as non-normal, and thus as informative, as possible. As a side-effect, ICA can identify components that are approximately sparse, showing many values around zero.

In this work, the FastICA algorithm by A. Hyvärinen [53] was used. As illustrated in Figure 1.3, the matrix A is split into the product $A = R\,C^{1/2}$, where the dewhitening matrix $C^{1/2}$ representing the linear correlations is calculated from the data covariance matrix C. The remaining rotation R is chosen such that the statistical dependence between the independent components is minimised. In order to avoid the time-consuming calculation of the H_k, FastICA substitutes the difference $H_N - H_k$ by a so-called contrast function

$$J_G(k) = |\langle G(S_{ik})\rangle_i - \langle G(\nu)\rangle|$$

J_G applies some even, non-quadratic function $G(\cdot)$ (in this work, the Gaussian function has been chosen) to each variable $S_{\cdot k}$ and to a normally distributed variable ν, returning the absolute difference of the mean values. In the algorithm, the matrix R is initialised with random values and then iteratively adjusted to maximise the J_G until a convergence criterion is met.

Like principal component analysis, ICA removes all linear correlations. To introduce a non-orthogonal basis, it also takes into account higher-order dependencies in the data. If the data lack such higher order structure, for instance, if they are normally distributed, the solution is not unique. The ICA model leaves some freedom to scale and sort the components: by convention, the independent components are scaled to unit variance, while their signs and their order can be chosen arbitrarily. The number of independent components equals the number of variables, but it may be reduced, for instance by removing weak principal components before applying the ICA, which considerably decreases the computational costs.

If the data fulfil the independence assumption (1.10), sparse components are easily found by ICA. Real data are probably noisy and the components are not exactly independent. There exist many variants of ICA with different assumptions about the distribution of the components. Noisy ICA explicitly takes into account additive noise. A supergaussian distribution for the mixing matrix can be easily implemented by modifying the data used in the FastICA algorithm [58]. Components estimated by ICA will still not be exactly independent. The remaining dependencies cannot be removed by a linear transformation, but they can be distributed differently between the components: independent subspace analysis [55] and topographic ICA [56] explicitly assume the existence of higher-order dependencies between the components, and try to distribute them among the components in a controlled way. Independent subspace analysis aims at concentrating the mutual information within subspaces of fixed dimensionality. With topographic ICA, a (usually grid-like) topology between the components is prescribed, and the components are

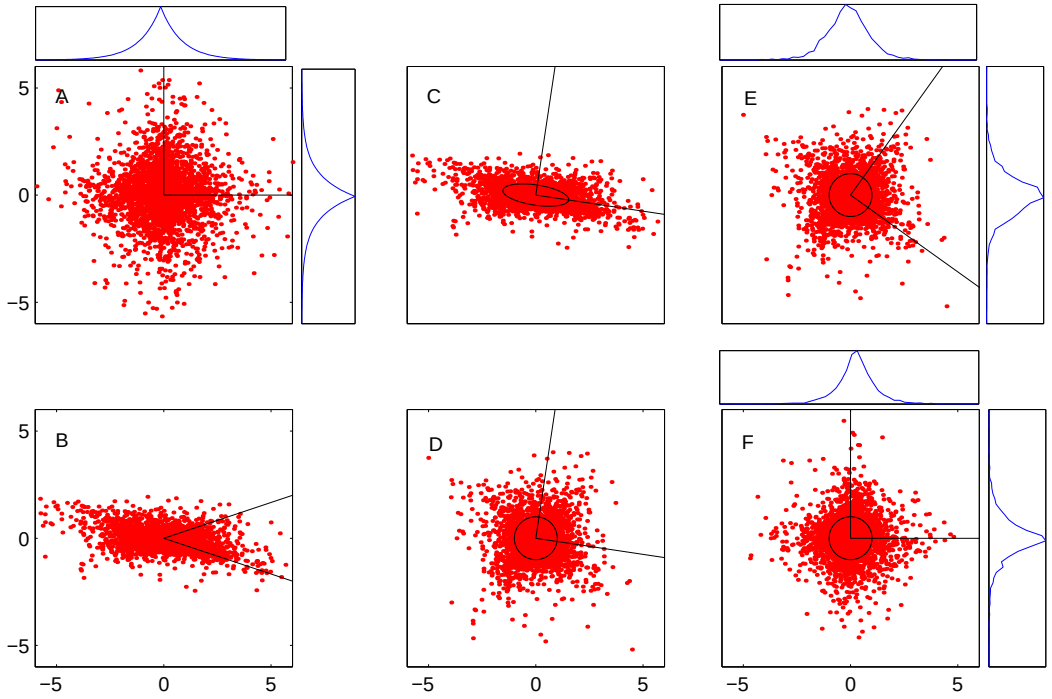


Figure 1.3: Independent component analysis of artificial data. Left: A cloud of n data points was produced by (A) choosing independent coordinates S_1 and S_2 from the two-sided exponential distribution and (B) shearing the data cloud by a linear transformation A. The centred data are contained in a $n \times 2$ matrix X. ICA reconstructs the unsheared data up to scaling, permutation, and reflection of the axes, based on the knowledge that the coordinates were independent. Centre: (C) Linear correlations between the two variables are represented by the covariance matrix C: its eigenvectors point along the axes of an ellipse defined by $\mathbf{x}\ C^{-1}\ \mathbf{x}^T = 1$. ICA "whitens" the data (D) by stretching them to unit variance along these directions, thereby removing the linear correlations. Right: The whitened data (E) are rotated to independent components (F) maximising the contrast function J_G, a dissimilarity between their marginal distribution and the normal distribution.

estimated such that the mutual information is concentrated between neighbouring components on the grid.

1.3.3 Other linear and nonlinear models

Different linear factor models have been applied to expression data by other authors or will be used in this work:

- **Factor analysis** is supposed to identify a few interpretable components ("factors") behind the data. The factors s_k are assumed to be independent Gaussian with unit covariance, and the noise variance is estimated separately for each variable. Factor subspace and noise term are separated by maximum likelihood, using the data correlation matrix. For detecting interpretable factors, it is postulated that the loadings matrix A must be almost sparse. This so-called "simple structure" is achieved by maximising some measure of non-Gaussianity: the varimax criterion, for instance, forces the sum of squared loadings to be maximal.

- With **overcomplete representations** [87], the number of components exceeds the number of data variables. As a maximum likelihood problem, it would be underdetermined, but it can be regularised by a prior distribution for the components.

- With **nonnegative matrix factorisation** (NNMF) [75], both data and components are constrained to be nonnegative. A (possibly overcomplete) decomposition according to maximum likelihood can be calculated by an iterative algorithm. NNMF tends to represent the data by sparse additive parts.

- **Correspondence analysis** searches for similarities between qualitative variables and has been applied to expression data in [28]. In this case, the data matrix was implicitly regarded as an contingency table, representing amounts of RNA molecules.

- The **plaid model** [74] explains expression data by processes that are specific for subsets of the genes and the samples. The data matrix is split into additive terms, each related to some of the rows and some of the columns of the data matrix, in analogy to biclusters. If each of these terms factorises into a gene-specific and a sample-specific part, the plaid model has the form (1.5) of a linear model.

- **Canonical analysis** studies whether two groups of statistical variables can be explained by common factors (called "canonical variables"). The variables of group 1 are linearly transformed to linearly uncorrelated canonical variables v_i, while the variables of group 2 yield canonical variables w_i, according to the following requirement: the first canonical variables v_1 and w_1 show a maximal linear correlation, the

second canonical variables v_2 and w_2 are maximally correlated under the constraint that they must be uncorrelated with their respective first canonical variable, and so on. For more than two groups, I shall use a generalisation of canonical analysis described in [100] in which the corresponding canonical variables for the different groups are obtained by projecting a single variable to the respective subspaces.

Other methods are based on nonlinear mixing functions (f, in equation 1.2): compared to linear models, nonlinear methods require more parameters to be fitted, the results are less well determined by the data, and the calculations can be quite expensive.

- **Self-organising maps (SOM)** [69] are clusterings where the clusters are arranged on a (usually cubic or hexagonal) grid. The cluster centres form a discrete coordinate system in the data space. Coregulation of genes has been visualised by SOM in [109]: the genes are distributed on a two-dimensional grid, such that coregulated genes are close to each other.

- With **nonlinear PCA** (see [22]), a neural network with a small number of nodes in the central layer is trained to map the data vectors to themselves. After training, the values in the central layer provide a low-dimensional representation of the data.

- **Nonlinear ICA** assumes statistically independent variables ("sources") behind the data, which are mapped to the observed variables by a nonlinear function that is represented by a neural network. The algorithm from [73] [72], which is used in this work, is based on Bayesian ensemble learning, that is, the joint posterior for all model parameters is approximated by a simplified parametrised distribution.

1.4 Mathematical cell models

To describe the structure and behaviour of biological systems, we must use concepts, expressed in words or mathematical terms, and thus create a model. In doing so, we implicitly choose a level of complexity for the description. Biological structures, as we see them, can be represented by mathematical structures.

1.4.1 Dynamical systems and genetic networks

A common mathematical framework to describe cells are dynamical systems, in particular systems of ordinary differential equations (ODE). ODE models can be used to simulate time series of the system. Bifurcation theory studies how different choices of the parameters influence the qualitative behaviour, such as stationary states and their stability, limit

cycles, or deterministic chaos. Cell models can be projected to smaller effective models with less variables. The reduced model may describe less details or an effective behaviour for certain time scales or under certain conditions. By the projection, variables may be lost, they may become parameters or be replaced by fewer effective variables, and new effective interactions between variables may arise. If an effective interaction cannot be derived mathematically (e.g. by time scale separation), it may simply be chosen to fit the observed data (or simulation runs from the full model) for the ensemble of conditions studied. In a detailed cell model, the interactions between different variables are sparse, while effective variables become probably more connected. For physiological states and certain experimental conditions, simple relations may hold between the cell variables: transients, attractors or distributions in the space of cells may be parametrised by a few global variables explaining most of the model dynamics: for instance, Hynne et al. [52] described oscillations of about 20 metabolites as linear combinations of two effective variables. These variables describe the systemic behaviour, namely small oscillations just above a Hopf bifurcation, without having a dedicated biological interpretation.

Mathematical models for metabolic and other cellular subsystems have been built based on biological knowledge and experimental data (see, for instance, [97] [11] [52]). The variables refer to the concentrations of important metabolites, while terms in the differential equations describe chemical reactions. Sometimes, though, small numbers of molecules have to be described by stochastic processes. Space-dependence is taken into account by compartments or partial differential equations. Currently, the main problem in large-scale cell modelling is the lack of detailed information about elementary processes, such as the kinetics of chemical reactions. For the regulation of gene expression, qualitative information can be deduced from the DNA sequence, for instance, about binding sites for transcription factors [43]. Studies like [116], where the program of a particular sea urchin gene was determined very accurately, are costly and will only be conducted for genes of special interest. Alternatively, genome-wide expression data can be used to build genetic network models, supposed to describe effective interactions between the genes. Quantitative models have been fitted to time series [19] [114], and sparse network topologies [99] [92] or influence strengths [17] have been determined by statistical analysis of expression data. Bayesian networks represent the dependence between statistical variables. Applied to expression data [31] [90], they explain the expression of each gene by the expression of a few "parent genes", but the dependencies described are purely statistical in nature, and need not correspond to biological interactions.

1.4.2 Metabolic control analysis

Metabolic control analysis (MCA) [41] [42] studies how stationary states of metabolic systems respond to changes of parameters. A metabolic system can be described by

differential equations for the metabolite concentrations

$$\dot{s} = Nv(s,p) \tag{1.11}$$

The vectors s, v, and p describe metabolite concentrations, reaction velocities (also called "fluxes"), and parameters (e.g., enzyme concentrations), respectively, while the stoichiometric matrix N contains the stoichiometric coefficients, each column describing one reaction. The behaviour of a metabolic system can be further characterised by the following quantities: K denotes a maximal kernel matrix of stationary fluxes, fulfilling $NK = 0$. The link matrix L [94] is defined by $N = L\ N^0$, where N^0 contains a maximal set of linearly independent rows of N, corresponding to a set of independent metabolites. The link matrix relates the concentrations of all metabolites to those of the independent ones and thereby describes the conservation relations. The reaction elasticities $\epsilon_{ik} = dv_i/ds_k$ describe the dependence of the reaction velocities on the metabolite concentrations in a linear approximation. Accordingly, the elasticities π_{ik} describe the linear influence of parameters p_k on the reaction rates v_i. The response coefficients R_E^S and R_E^J describe the linear influence of parameters (in this case: enzyme concentrations E_k) on steady state concentrations S and fluxes J, and they can be decomposed into a product $R_E^J = C^J\ \pi_E$ (similar for R_E^S). The control coefficients C^J and C^S describe the change of steady-state concentrations or fluxes due to a small parameter change affecting only one reaction

$$(C^J)_k^i = \frac{\partial J_i/\partial p}{\partial v_k/\partial p} \tag{1.12}$$

$$(C^S)_k^i = \frac{\partial S_i/\partial p}{\partial v_k/\partial p} \tag{1.13}$$

and can be calculated by (see [41])

$$C^S = -L(M^0)^{-1}N^0 \qquad \text{with the Jacobian} \quad M^0 = N^0\epsilon L \tag{1.14}$$

$$C^J = 1 + \epsilon C^S \tag{1.15}$$

They fulfil the summation and connectivity theorems of metabolic control theory

$$\begin{pmatrix} C^J \\ C^S \end{pmatrix} (K \quad \epsilon L) = \begin{pmatrix} K & 0 \\ 0 & -L \end{pmatrix} \tag{1.16}$$

To deal with logarithmic values of concentrations, fluxes, and perturbation parameters, the control coefficients are replaced by the normalised control coefficients $\mathrm{dg}(S)^{-1}\ C^S \mathrm{dg}(J)^{-1}$ and $\mathrm{dg}(J)^{-1}\ C^J \mathrm{dg}(J)^{-1}$, respectively.

Metabolic control coefficients are related to projection operators[2] in the space of flux distributions [94], as shown in Figure 1.4. The columns of the kernel matrix K span the

[2] A projector is a linear operator fulfilling $P = P^2$.

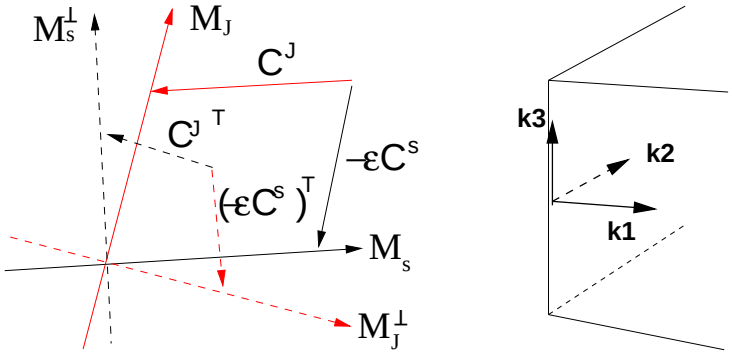


Figure 1.4: Space of the metabolic fluxes. Left: Metabolic control coefficients are related to projection operators in the space of flux distributions. The space of all flux distributions is spanned by the subspaces $M_J = \text{span}(K)$ and $M_S = \text{span}(\epsilon L)$. According to the theorems 1.16, the matrix C^J of flux control coefficients acts as a projector to M_J, while $-\epsilon C^S$ projects to M_S. The transposed matrices C^{J^T} and $(-\epsilon C^S)^T$ project to the perpendicular subspaces $M_S^\perp \perp M_S$ and $M_J^\perp \perp M_J$. Right: Elementary flux modes. The irreversible flux modes k_1 and k_2 and the reversible mode k_3 (shown as vectors) span the convex space of admissible flux distributions. Any stationary flux distribution can be decomposed into a linear superposition of the elementary modes, in which irreversible modes appear with nonnegative coefficients.

space M_J of steady-state fluxes, while the columns of ϵL span the space M_S of immediate flux changes resulting from small virtual changes of the independent concentrations (compare Figure 6.2). The spaces M_J and M_S are linearly independent, but in general not orthogonal, and span the space of all flux distributions. The summation and connectivity theorems 1.16 imply that the matrices C^J and $-\epsilon C^S$ project to M_J and M_S and sum to the identity matrix. Accordingly, the transposed matrices $C_x^{J^T}$ and $(-\epsilon C_x^S)^T$ project to the respective perpendicular subspaces $M_S^\perp \perp M_S$ and $M_J^\perp \perp M_J$.

Any stationary flux distribution in a metabolic network can be decomposed into a superposition of elementary flux modes [104], which can be interpreted as metabolic pathways. For calculating the elementary modes, the stoichiometric matrix and knowledge about the reversibility of reactions are required. Some of the elementary modes, called "irreversible modes", can only appear with nonnegative coefficients in the superpositions. Geometrically, the elementary modes span the space of admissible flux distributions, and the subspace spanned by the irreversible modes is a convex cone (Figure 1.4, right). Each face of the admissible region is spanned by all reversible and some of the irreversible modes. The elementary modes are unique, but may be overcomplete, so the decomposition of a flux distribution into elementary modes is possibly not unique. Elementary modes can also be used for describing optimal flux distributions: if normalised flux distributions are rated by a linear fitness function, then the optimum probably lies on a face of the cone, represented by few elementary modes.

1.4.3 Mathematical notions of information

If a biochemical event is called a "signal", it is supposed to carry information. Two mathematical concepts of information are relevant for this work, namely the mutual information and the information value.

The **mutual information** $I(X, Y)$ [16] characterises the statistical dependence between two random variables X and Y. It is defined as

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \tag{1.17}$$

where H denotes the Shannon entropy. The Shannnon entropy for a discrete random variable with values x_i is defined as

$$H(X) \;=\; -\sum_i p(x_i) \log_2 p(x_i) \tag{1.18}$$

For continuous variables, the sum is replaced by an integral. If X and Y are independent, i.e., if $p(x, y) = p(x)\ p(y)$, then the mutual information between them vanishes, while otherwise, it is positive. Thus if the mutual information is high, then knowing a value

of X allows for a better prediction of Y, and vice versa. If there is a large mutual information between a biochemical signal X and some other process Y, a third process can respond to X as if it received a direct input from Y. The process Y may be located in the signalling pathway upstream of X, but by statistical correlation, signals can also provide information about processes in the future.

Even if the mutual information between a signal and some other process is high, the information provided may be irrelevant for the cell. Relevant information can be quantified by the **value of information** defined in Bayesian decision theory (see [89]). Let us assume somebody who receives noisy signals from its environment and then has to choose between different possible actions. The actions are rated by a pay-off which depends on the action chosen and on the actual state of the environment. The value of an information source is defined as the difference between the expected pay-offs gained with and without using the signals. If use of the information source causes additional costs, it should only be used if the information value exceeds the costs. The value of an information source may strongly depend on the other information sources present. High mutual information and information value of signals need not coincide unless a signal carries mutual information about (otherwise unknown) conditions that need to be known for a good decision. The concept of information value is actually supposed to describe intelligent systems: however, if biochemical signal processing systems have evolved to behave optimally, they should behave like an equally informed, intelligent system, and thus follow the rules of Bayesian decision theory.

Part I

Analysis of expression data

Chapter 2

Gene programs and expression modes

This chapter is concerned with a mathematical description of gene expression. Each gene is characterised by a "gene program", a mathematical function stating how expression depends on cellular variables called "expression modes". A model for simulating expression data is proposed, and gene programs are estimated by regression from simulated and real gene expression data.

2.1 A mathematical description of gene expression

2.1.1 Gene programs and expression modes

The expression of a gene can be modelled, on a molecular level, by a stochastic process describing synthesis and decay of mRNA molecules. A stochastic model is only necessary for rare transcripts, while large mRNA numbers are sufficiently described by a deterministic model which represents the ensemble- or time-averaged behaviour of the underlying stochastic process. A time-dependent mRNA concentration $x(t)$ then follows the differential equation

$$\dot{x} = \sigma(y_\sigma) - \mu(y_\mu)\, x \tag{2.1}$$

where the synthesis rate σ and the decay rate constant μ depend on vectors y_σ and y_μ of input variables (see Figure 2.1). The inputs y_σ may represent, for instance, the local concentrations of transcription factors in their active form. If a gene is controlled by a signalling chain, the choice of the input signals signal is somewhat arbitrary (see Figure

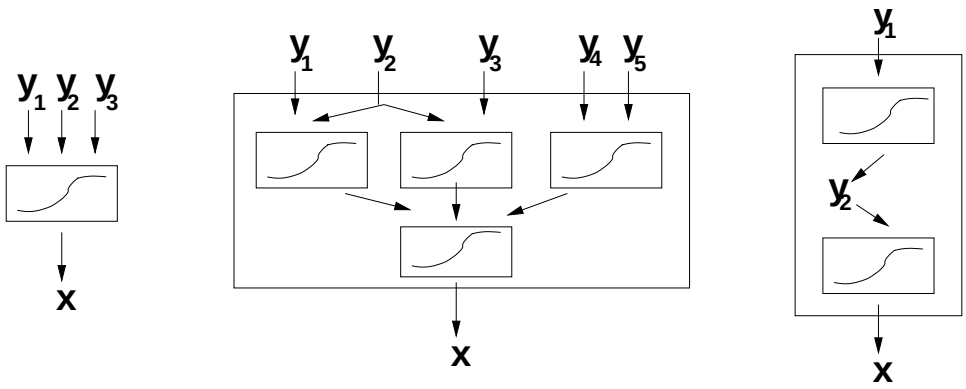


Figure 2.1: Gene programs. Left: The expression value x of a gene is described by a nonlinear function (called the "gene program" and shown as a box), which depends on input variables y_i. The expression value can represent the synthesis rate σ or alternatively, the steady-state concentration $x^{(stat)}$ of mRNA. Centre: Nonlinear functions (small boxes) can be combined: in (artificial) neural networks, such combinatorial schemes are used to parametrise arbitrary nonlinear functions. In gene programs, the combination of regulatory functions may represent the combinatorics of transcription factors forming complexes or binding to different sequence motifs on the DNA. Right: A gene is regulated by a cascade of regulators y_1 and y_2. A program for the gene x can be defined with respect to either of them. If the upstream signal y_1 is regarded as the input, the biological mechanism involving y_2 becomes part of the gene program.

2.1, right): the synthesis rate could also be written, effectively, as a function of more distant signals. Nevertheless, each gene is directly influenced only by a few signals, so in the cell model, the direct influences are sparse. Transcription factors can act on many genes, so there may be considerable overlap between the input signals of different genes.

The general solution of equation 2.1 for given time courses $x(t_0), y_\sigma(t)$, and $y_\mu(t)$ is

$$x(t) = x(t_0)\ e^{-\int_{t_0}^{t} \mu(y_\mu(\tau))\mathrm{d}\tau} + \int_{t_0}^{t} \sigma(y_\sigma(\tau))\ e^{-\int_{\tau}^{t} \mu(y_\mu(\tau_1))\ \mathrm{d}\tau_1}\ \mathrm{d}\tau \tag{2.2}$$

For $t_0 \to -\infty$ and constant decay rate $\mu > 0$, this becomes

$$x(t) = \int_{-\infty}^{t} \sigma(y_\sigma(\tau))\ e^{-\mu\cdot(t-\tau)}\ \mathrm{d}\tau \tag{2.3}$$

The stationary state solution for constant input signals y_σ and y_μ is

$$x^{(stat.)} = \frac{\sigma(y_\sigma)}{\mu(y_\mu)} = f(y_\sigma, y_\mu) \tag{2.4}$$

Thus either the synthesis and decay rates or the stationary expression level itself are described by a nonlinear function of several input variables (see Figure 2.1). In the

following, I shall refer to the concentration of mRNA rather than to its synthesis rate: however, both quantities are closely related if mRNA degradation is not controlled and if the cell processes studied are much slower than the time scale of mRNA decay (tens of minutes).

Gene expression need not be explained by actual biological signals: if cell signals are correlated with some global variable in the ensemble of conditions studied, this global variable may be used as an effective input signal. Generally, I shall suppose that an expression value $x^{(i)}$ can be described by the "gene program", a function

$$x^{(i)} = f^{(i)}(y) \tag{2.5}$$

of variables y called "expression modes", which represent either actual biological signals or global variables.

2.1.2 Simulated gene expression data

Artificial data are useful as a benchmark for analysis methods. In this section, I shall propose a model for simulating gene expression according to the ideas mentioned above: the cell state is characterised by unobserved variables y_k called "expression modes" to which each gene $x^{(i)}$ responds according to a nonlinear gene program $f^{(i)}(y)$ given by

$$f^{(i)}(y) = x_\infty^{(i)}\, g(-w_{i0} + \sum_k w_{ik}\, y_k) \tag{2.6}$$

$$\text{with} \quad g(z) = \frac{1}{e^{-4z}+1} \tag{2.7}$$

The functional form (2.6) is widely used for defining nonlinear functions with few parameters (compare, e.g., [114]). The sigmoidal activation function $g(z)$ saturates at 0 and 1 for small and large values of z, respectively. It is antisymmetric, fulfilling $g(-z) = 1 - g(z)$, and has a slope of 1 at $z = 0$. The nonlinearity evokes an interaction between the input signals y but for small arguments of g, the model is approximately linear, and the inputs weights w_{ik} can be compared to the parameters of linear models. Equation 2.6 can only describe functions with plane isosurfaces, that is, straight contourlines in the two-dimensional case. This is a strong restriction, but the assumption is supported by the relatively weak nonlinearities found behind experimental data in chapter 3.

To simulate expression data, the model parameters $x_\infty^{(i)}, w_{i0}, w_{ik}$, as well as the noise terms are drawn independently from random distributions. The maximal values $x_\infty^{(i)}$ and offsets w_{i0} are log-normal and normal, respectively. The input weights w_{ik} are sparse, so genes receive signals only from some of the modes, and the non-vanishing weights are independent normal. Sparsity is controlled by the mean number of inputs per gene.

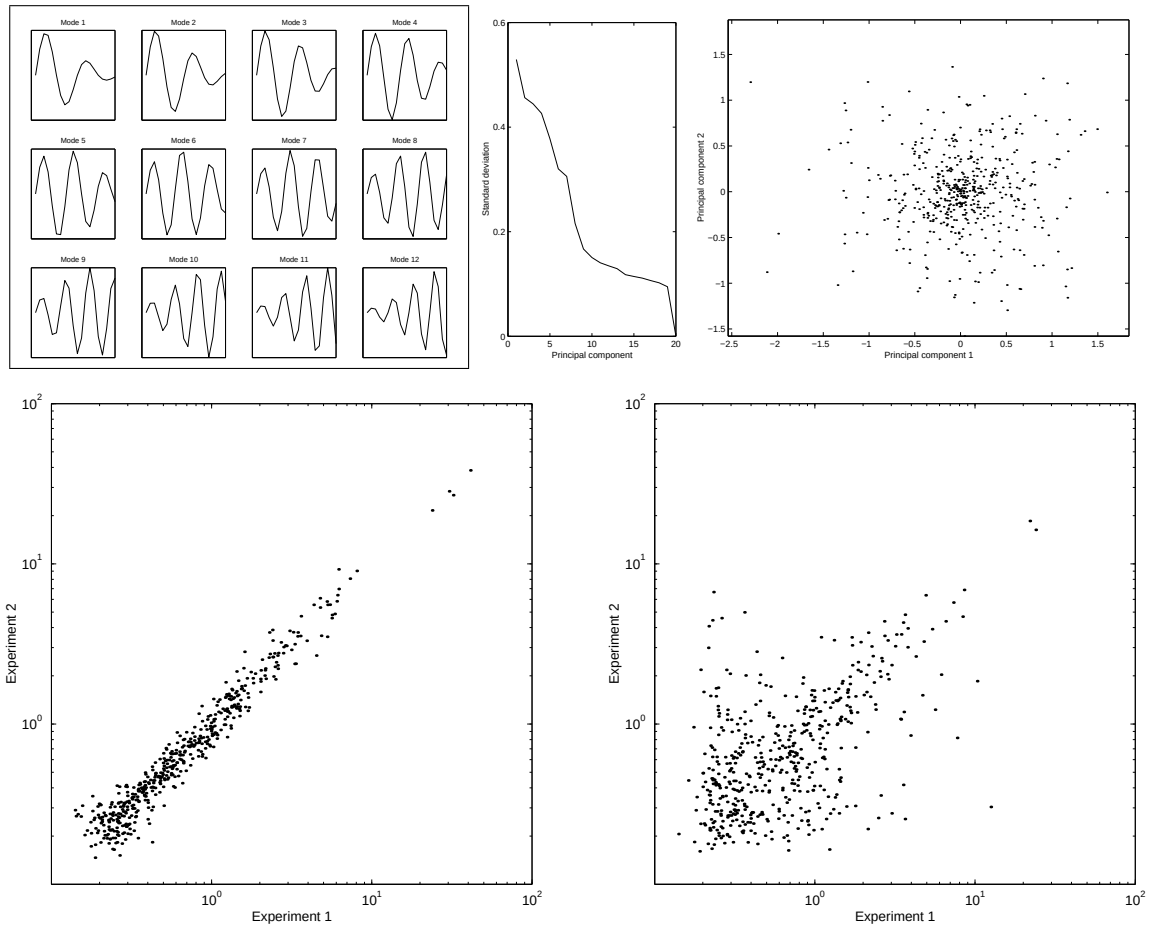


Figure 2.2: Simulated gene expression data. Expression values x for 50 time points and 500 genes were simulated using equation 2.6. Twelve expression modes y_k were used as inputs for the gene programs, and each gene responds, on average, to 4 of them. Noise was modelled by additive and multiplicative terms. The model parameters are listed in the Appendix B.1. Top left: As time series of the twelve modes, modulated sinus waves were chosen. Top centre and right: Principal components of the simulated data. The diagrams show the standard deviations of the principal components and a scatter-plot between the first two principal components. Each point represents a gene profile. Bottom: Scatter-plots between two simulated experimental samples (microarrays). Left: By varying only the noise terms, a repeated experiment was simulated. Right: Both the expression modes and noise were varied to model different experimental samples. Points far from the main diagonal represent differentially expressed genes.

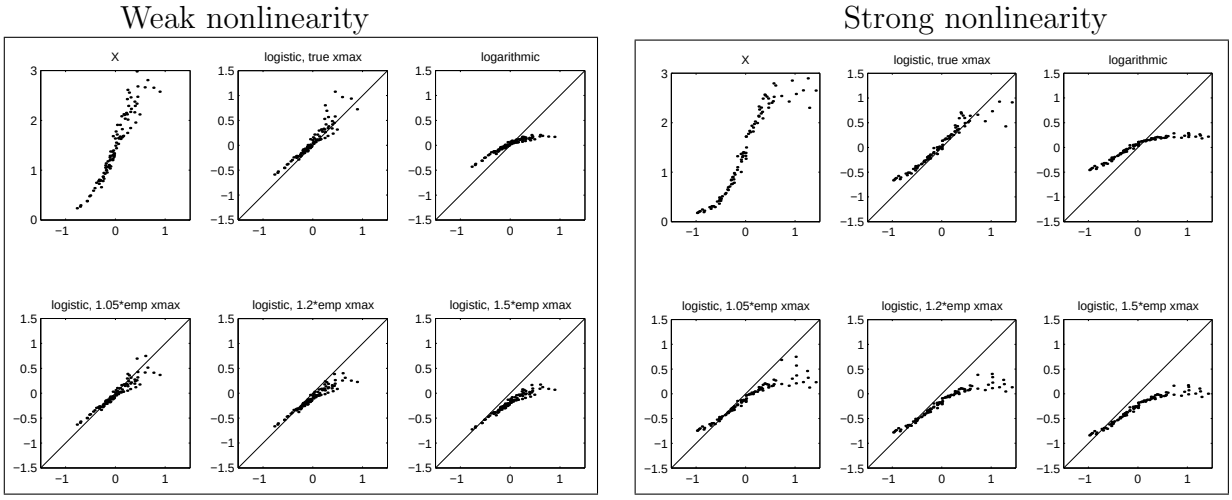


Figure 2.3: Preprocessing the data can reduce nonlinearities. Different transformations were applied to simulated expression data to reconstruct the linear activation (argument of the nonlinearity $g(z)$ in equation 2.6). The large boxes show results for different choices of the simulation parameters (see appendix, table B.1, parameter set (1)), yielding an almost linear gene program (left box, with normally-distributed $w = 0 \pm 0.2$) and a rather nonlinear gene program (right box, $w = 0 \pm 0.4$). Results for other parameter choices are shown in Figure C.2 in the appendix. The small diagrams refer to different transformations. Top left: Artificial data for a gene. The expression value x is plotted against the linear activation z. The other five boxes show the same data, with different transformations applied to x. The diagonal line indicates the true activation values to be reconstructed. Top centre: Reconstruction with the logistic transformation (2.8) and the true x_∞. Top right: Log-transformation (natural logarithm). Bottom row: Logistic transformation, with x_∞ guessed by $\gamma \max(x)$, for $\gamma = 1.05, 1.2, 1.5$. The reconstruction by the logistic transformation is more accurate than the reconstruction by log-transformation.

Finally, some of the genes are supposed to have no inputs at all, showing only random fluctuations. For illustration, the expression modes are modelled by modulated sinusoidal time series. The noise is modelled by additive and multiplicative terms, both log-normal and chosen independently for each data point. Simulated data for 50 experiments and 500 genes are shown in Figure 2.2. The parameter sets for all simulations used for this text are listed in Appendix B.1.

2.1.3 Reducing the saturation effects in expression data

In analyses of gene expression data, it is common to use log-transformed data, because they often show a simpler noise distribution. In addition, it has been argued [9] that lin-

ear factor models, applied to log-transformed data, describe multiplicative effects behind the untransformed data, which is biologically plausible if the combinatorial interaction of different transcription factors is considered multiplicative rather than additive. Thus combined with a proper preprocessing, a linear model can describe nonlinear combinatorial control. The logistic model (equation 2.6), which additionally describes saturation, suggests a different way to preprocess the data: the data $x = g(z)$ are transformed by

$$z = \frac{1}{4} \log \frac{x/x_\infty^{(i)}}{1 - x/x_\infty^{(i)}} \tag{2.8}$$

to the activation signals z which can then be split into their linear contributions. The maximal value x_∞ needs to be known to locate the onset of saturation: it cannot be easily estimated from the data, so practically, it must be guessed. Nevertheless, if real expression data contain strong saturation effects, this logistic transformation may still be more appropriate than the log-transformation. To test this, I applied both transformations to noisy artificial data, trying to reconstruct the activation values $z = -w_{i0} + \sum_k w_{ik} y_k$. The value x_∞ was guessed for each gene separately by $\gamma \max(x)$, with different values of the factor γ. Results for simulated data with different parameter choices are shown in the Figures 2.3 and in Figure C.2. For the conditions studied, the logistic transformation, even with rough guesses of x_∞, outperforms the log-transformation. Nevertheless, the log-transformation will be used throughout this thesis in order to make the analyses comparable to those from other studies.

2.2 Estimation of gene programs

2.2.1 Estimated gene programs and expression modes

The scheme shown in Figure 2.4 explains the coregulation of genes by shared input signals: if the modes represent transcription factors, then the connections between modes and genes may represent corresponding binding sites. Accordingly, expression data have been used to determine motifs in the regulatory sequence [6] [9]. Here, on the contrary, models of the same form will be determined from the expression data alone. The statistical components behind the data and the corresponding loadings are interpreted as empirical gene programs and empirical expression modes. The variation of the modes between the samples is not further explained by a dynamical model or direct interactions between the genes. Statistical expression modes y_k and linear influence weights w_{ik} are estimated blindly from the data. An aim of this thesis is to study the biological significance of such estimated modes: even if they separate variation caused by biological processes from experimental artefacts, the separate modes may not have an obvious interpretation,

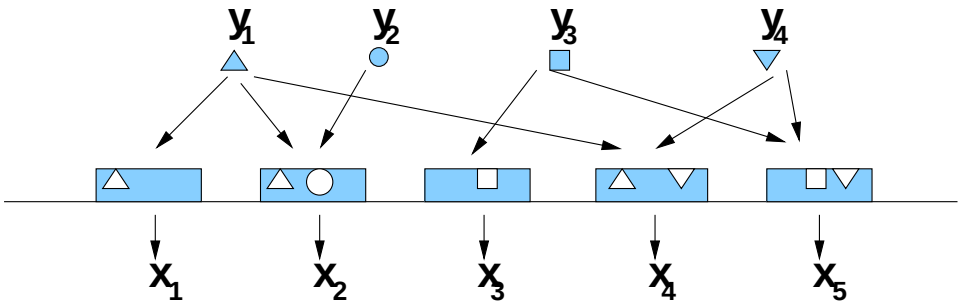


Figure 2.4: Gene programs and their inputs. Expression of the genes x_i is explained by common input variables ("expression modes") y_k. If the modes represent transcription factors, connections between modes and genes may indicate the presence of the corresponding binding sites (symbolised by boxes) on the DNA. In the mathematical model 2.6, gene programs are parametrised by influence weights w_{ik}, characterising the linear influences of the different modes. If each mode influences only some of the genes, then the influence weights are sparse. If expression modes represent global cell properties, they will probably influence all genes, but the input weights may still be almost sparse, with many weights close to zero. Models of the same form can be estimated blindly from expression data: in the statistical factor models studied below, many genes x_i are explained by a few empirical expression modes y_k.

because only few modes can be extracted from noisy data, so different biological processes will possibly not appear separately, but as a mixture.

Although the biological mechanisms behind expression involve many regulatory signals and processes, most of them are probably not visible in the data because their variation between the experiments is too weak to be silhouetted against the noise. Nevertheless, models of the form 2.5 can be reconstructed from expression data if many genes share their input signals and these signals vary sufficiently between the samples. For blind estimation, statistical assumptions are made about the components, in particular sparsity and statistical dependencies between them. The connections to different input signals may be statistically dependent [27]. If the modes represent effective global variables, then the inputs are probably less sparse, and genes involved in similar processes are likely to share inputs.

2.2.2 Biological interpretations of expression modes

The statistical properties of the expression modes and of the corresponding gene programs depend on their interpretation. In the following, I shall discuss alternative interpretations for the biological modes, namely a causal, a systemic, and a functional one. Besides this, also experimental artefacts my appear as modes.

1. **Causal**: In detailed cell models, expression depends on biological signals carried by molecules, such as transcription factors, MAP kinases, or membrane receptors. A gene program describes how expression is influenced by these signals. The number of modes is large, and the connections between modes and genes are sparse. The statistics of signals and gene programs depend on the kind of signals considered, and on the ensemble of cell states studied. The statistical dependencies between the input weights represent the structure of the signalling system: if genes share transcription factor binding sites or if the signalling chains are cross-linked, their input weights are dependent. A weak dependence between the input signals can be claimed based on an information-theoretical argument: for a fixed number of signalling channels, maximising the amount of transmitted information would require that the mutual information is minimised, so the signal values must be statistically independent. However, redundant signals may even be preferable because they are more robust against perturbations.

2. **Systemic**: As the gene expression machinery is part of a regulatory feedback loop, the distinction between processes upstream and downstream of gene expression is slightly artificial. In an ensemble of experimental conditions, the cell state, including gene expression, may be characterised by relatively few global variables, and maybe, only a few of them are on the same time scale as expression. The separate global variables need not have a simple biological meaning: they are simply chosen to parametrise the cell state. The statistical dependencies between them should be weak, for conceptual reasons, because otherwise, a different choice of global variables would be more appropriate.

3. **Functional**: In the teleological framework of chapter 5, the gene programs describe the function of genes rather than their regulatory mechanism, so gene expression will be explained by cell variables for which the gene is responsible. Necessary changes of these variables will play the role of expression modes. It will be shown that the optimal input weights w_{ik} reflect the response coefficients and the fitness curvatures. Simulations of the control coefficients show that these input weight should be almost sparse and related to the topology of the metabolic network. Accordingly, programs of functionally related genes should be statistically dependent. This interpretations does not necessarily contradict the causal one: if the regulatory system is optimised, the causal inputs may reflect the function of genes, as we shall see in section 5.2.2.

4. **Artefacts**: Also artefacts from the sample preparation or hybridisation procedure may appear as empirical modes, representing, for instance, genes with special hybridisation properties. If these hybridisation properties are independent of the biological roles of the genes, ICA may be able to separate both kinds of effects.

2.2.3 Reconstructing a gene program behind artificial data

Before gene programs will be estimated blindly in the following chapters, I shall study first, as a test, how they can be reconstructed if both input signals and expression values are known: regression is applied to artifical data to see how a linear model performs on nonlinear data and to study the effect of noise. The model from section 2.1.2 was used to simulate the expression of a gene under the control of three expression modes in 100 experiments. A linear and a nonlinear regression model were fitted to the noisy data, both in their original form and after log-transformation. In Figure 2.5, model fits and predictions (calculated by cross-validation) are plotted versus the data with and without noise. For the cross-validation, the samples were divided into four equally-sized groups, and the values for each group were predicted by the model fitted to the data from the other groups.

Linear regression overestimated the extreme values because it could not handle the nonlinearity behind the data. Nonlinear regression using a two-layer perceptron did not suffer from this problem. Apart from this weakness of the linear model, the model fits are quite satisfactory. Due to the additive noise term, there is an offset between the clean and the noisy data, becoming particularly prominent after log-transformation. Except for this offset, the predictions from the nonlinear model, without log-transformation, fit the clean data better than the noisy ones although the model was trained with the noisy data. After the results of cross-validation, this indicates again that no over-fitting occurred.

2.2.4 Explanatory variables for gene expression

Transcription factors might be good explanatory variables for modelling gene expression. On the other hand, global variables describing the cell state may also allow for predicting the expression of single genes: both approaches are compared in this section. For the calculations, the 1000 most variant genes were chosen from the stress response data set Causton et al. [10] and the 100 genes to be explained we randomly chosen from these 1000. Transcription factors that respond to themselves were discarded. From the remaining 900 genes, principal and independent components were calculated (as described below, in section 4.1), and their expression modes were used as candidate explanatory variables. In [76], binding of transcription factors to the regulatory regions of yeast genes was studied experimentally with microarrays and quantified by p-values: independent components from these binding data are shown in Figure 2.6. The binding data were used to choose, for each gene, the n most probable transcription factors. As the concentrations of the transcription factors had not been measured, their own expression values, as a measure of their activities, were also used as explanatory variables.

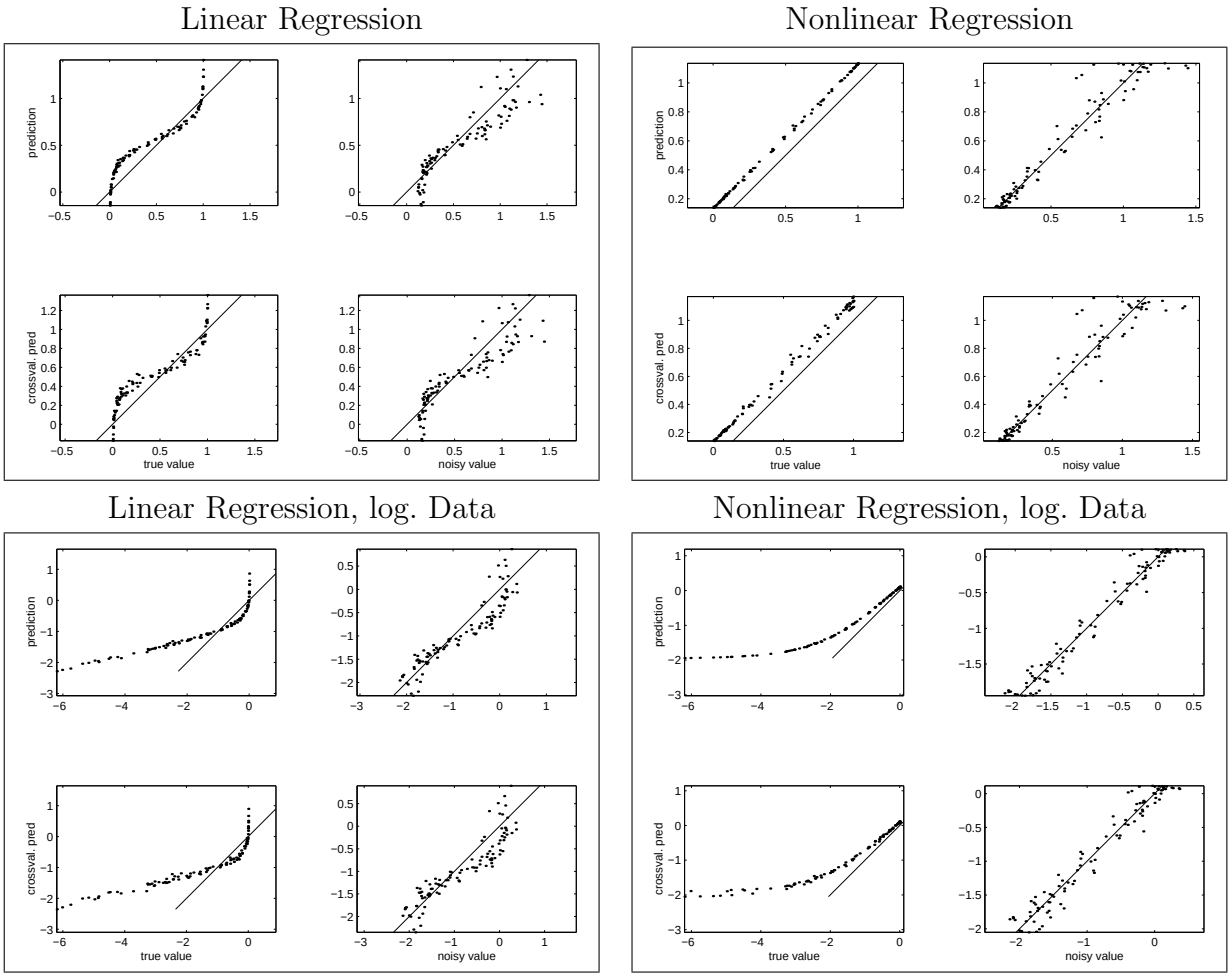


Figure 2.5: Regression with artificial data. Left: Noisy expression data were simulated for a gene x with 3 inputs y_k, for 100 time points (parameters see appendix, Table B.1 (2)). The nonlinear function $x(y_1, y_2, y_3)$ used for the simulation was estimated from the data by linear and nonlinear regression, using the y_k as explanatory variables. The large boxes show results for original (top) and log-transformed (bottom) data, and for linear regression (left) and nonlinear regression using a two-layer perceptron (right). Inside each large box, predictions are plotted versus the data. The top and bottom plots show model fits and predictions from cross-validation, respectively. The plots on the left and on the right show clean and noisy original data - for the regression, only the noise data have been used. Due to the additive noise term, there is an offset between true and noisy data.

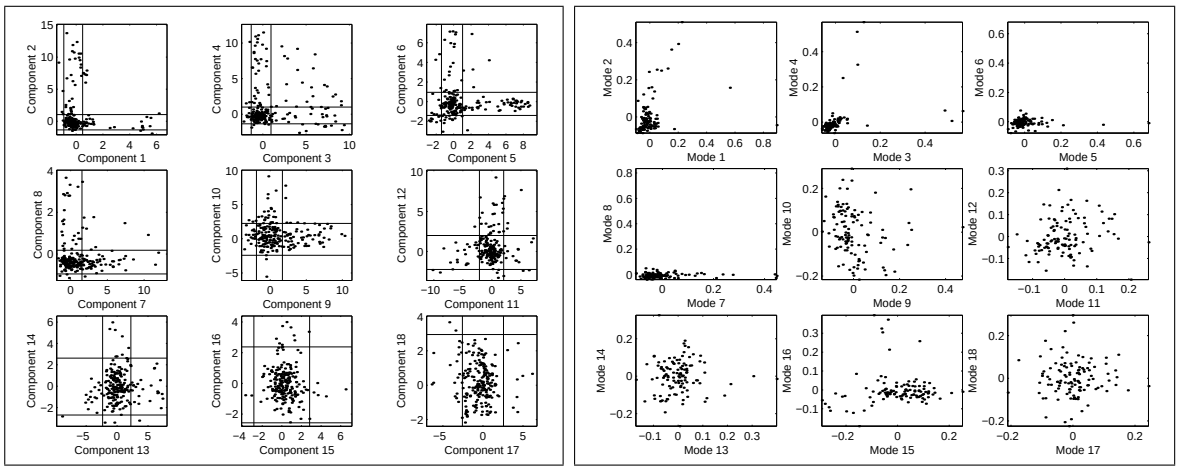


Figure 2.6: Independent component analysis of transcription factor binding data [76]. The data matrix X contains logarithmic ($\log_{10}$) p-values for binding of transcription factors (matrix columns) to the regulatory regions of genes (matrix rows). By ICA, X was decomposed into a product $S\ A$. The diagrams show scatter-plots between subsequent components (columns of S, left) and loadings (rows of A, right). Each of the first components defines a specific group of responding genes and corresponds to some of the transcription factors. The components may correspond to regulatory sequence elements, but this has not been tested here.

Linear and nonlinear regression were done for each of the 100 genes, and for each kind of explanatory variables. The root-mean-square prediction error, calculated by cross-validation with four groups, was normalised by the standard deviation of the respective gene. Figure 2.7 shows the distributions of these relative errors, for $n = 1, .., 10$ inputs per gene and for linear (top) and nonlinear (bottom) regression. The bars indicate the median and the quartiles over the 100 genes, each bar representing a type of explanatory variables: (1) the n first PCA modes, (2) the n first ICA modes, (3) the expression of the n transcription factors supposed to be most important for the gene, and (4) of n randomly chosen genes. Although the quality of the predictions varies largely among the genes, the distributions of the prediction errors show differences among the methods: for few inputs, the global modes from PCA and ICA yield better predictions than the other, individually chosen variables, while for larger n, the difference decreases. Astonishingly, the transcription factors do not seem to carry more information than genes chosen at random. So possibly, the transcription factors chosen based on binding affinities are not the most relevant regulators, or the expression of the transcription factors is maybe no reliable measure for their activity.

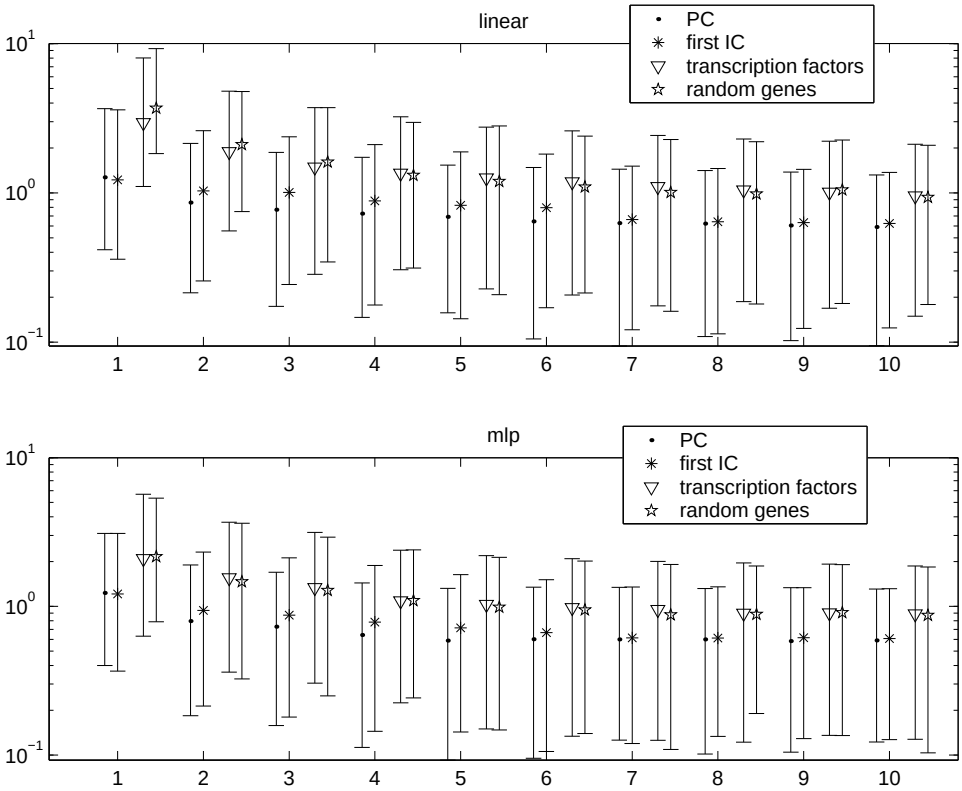


Figure 2.7: Estimation of gene programs from the cell stress data set Causton et al. [10]. The diagrams show prediction errors for experimental expression data, using different regression models and kinds of explanatory variables. Top: Expression of 100 genes was fitted by linear regression. For each gene, the following kinds of explanatory variables were considered: the n first PCA modes (dot), the n first ICA modes (star), the expression of the n transcription factors supposed to be most important for the gene (triangle), and of n randomly chosen genes (pentagram). The diagram shows the relative prediction errors from cross-validation, for different numbers n (abscissa) and kinds (shown by symbols) of explanatory variables. Each bar indicates the median and the quartiles of the relative prediction error over 100 genes. Bottom: The same, for nonlinear regression with a two-layer perceptron. For small numbers of explanatory variables, the global variables determined by PCA and ICA yield better predictions than the expression of the transcription factors.

Chapter 3

Estimating gene programs by nonlinear ICA

In this chapter, nonlinear gene programs are determined blindly from expression data by nonlinear ICA [65]. The genes are regarded as the statistical variables, and expression is explained by a few statistically independent modes called the "sources".

3.1 Test with simulated expression data

For the calculations, the algorithm described in [73] was used, with 8 hidden neurons in the neural network, and 5000 iterations per run. The algorithm uses ensemble learning to fit a posterior distribution of the model parameters. Here, I shall only refer to the "mean" model at the centre of mass of the posterior. For the estimation, the number of samples should exceed the number of genes, so a few genes were chosen from genome-wide microarray data. A variant of cross-validation was used to detect over-fitting. Before studying experimental data, I tested whether nonlinear ICA could reconstruct the model behind simulated data (see Figure 3.1). Figure 3.2 shows the model fits from nonlinear ICA for twelve genes: although the model was trained with noisy data, it represents the clean data better than the noisy ones, as the noise is partly averaged out in the estimation.

3.2 Application to experimental data

Nonlinear ICA was applied to the 50 most variant genes in the stress response data set [32], with 1, 2, or 3 source variables, which will be interpreted as expression modes. In the experiments, time series of expression in yeast had been measured under different stress

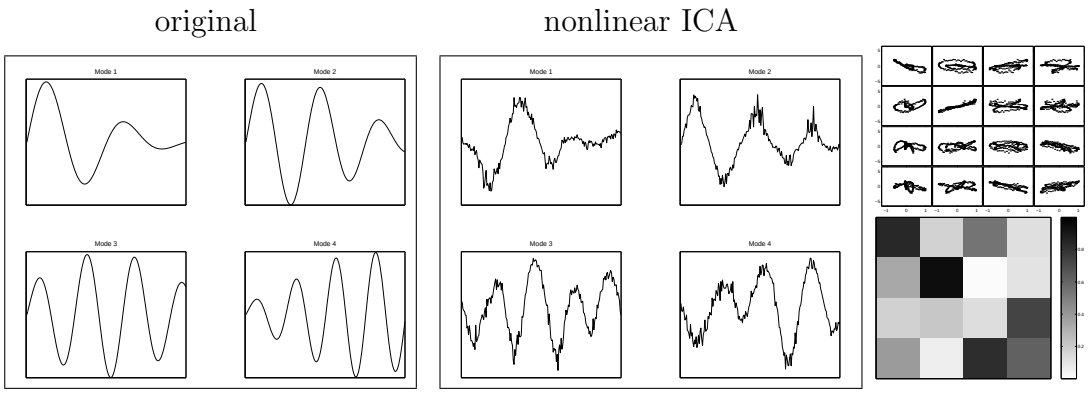


Figure 3.1: Nonlinear ICA applied to simulated expression data. Artificial data for 50 genes were produced using 4 expression modes y_k (parameters see appendix, Table B.1 (3)). Left: Time series used for the modes. Centre: Nonlinear ICA approximately identified them, except for their order (1,2,4,3) and the signs $(-1, 1, -1, -1)$. Right: The small diagrams show scatter plots and the matrix of absolute correlation coefficients between true modes (ordinate) and estimated modes (abscissa).

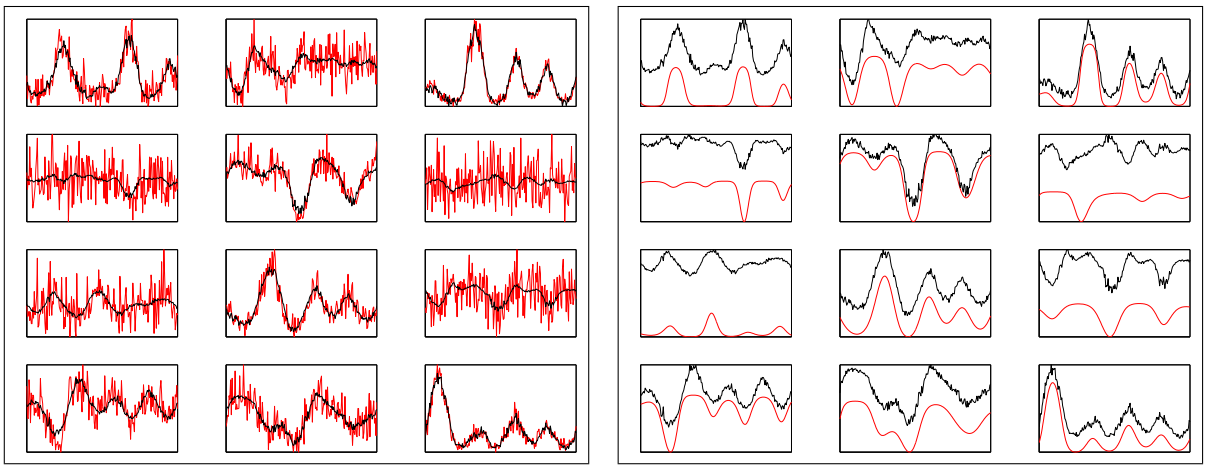


Figure 3.2: Model fits from nonlinear ICA. The diagrams show simulated time series of twelve out of 50 genes (compare Figure 3.1) along with their fits by nonlinear ICA. The red curves show the original values, with (left) and without noise (right), while the fitted curves are shown in black. Although the model was trained with the noisy data, the fits closely match the clean data except for an offset due to additive noise (compare the regression shown in Figure 2.5).

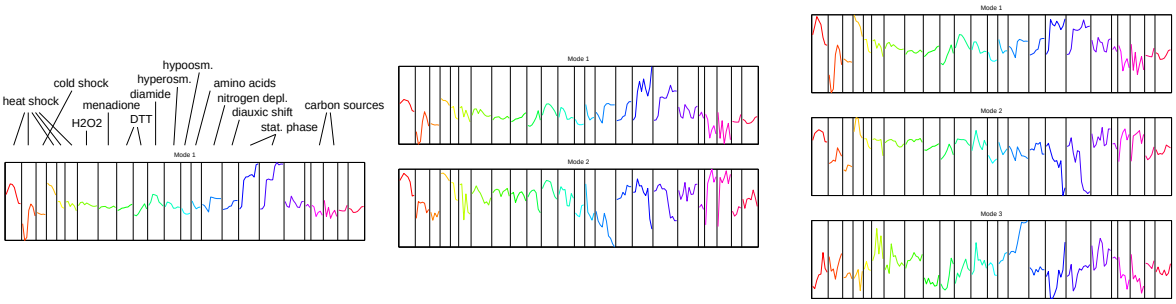


Figure 3.3: Expression modes behind the stress response data Gasch et al. [32], determined by nonlinear ICA. Time courses of 1 (left), 2 (centre), and 3 (right) estimated modes are shown. As the number of modes increases, the previous modes remain almost unchanged, and a new mode appears.

conditions, including heat shock, cold shock, oxidative stress, and hyper- and hypoosmotic shock, and the treatment with the sulfhydryl-oxidising and disulfide-reducing agents diamide and DTT. In addition, growth on different media (amino acid starvation, nitrogen source depletion), and the progression into the stationary phase was studied. Figures 3.3 and 3.4 show the time series of the estimated modes and the nonlinear programs of 12 genes, respectively. In most cases for one mode (Figure 3.4 top left), the nonlinearity has approximately a sigmoidal shape. For two modes (top right) the shapes become more complicated: while for gene HSP26, the contour lines of the nonlinear function are approximately straight lines, JEN and YGR052W show show a combinatorial control. In the periphery, though, the fits depend on a few data points and are probably not well determined by the data.

Fits for twelve genes are shown in Figure 3.5, left. Although the numbers of parameters (640, 822, 1004, for 1, 2, or 3 modes) are much smaller than the number of data points (8700), the results might still suffer from over-fitting. This was tested by a variant of cross-validation (see Figure 3.5). Both samples and genes were split into training and test sets of equal size, and nonlinear ICA was applied to the training set of samples, yielding programs for all genes. Then, a linear model was fitted to explain the modes by a linear combination of the training genes, so for the test samples, the modes could be predicted from the training genes. Finally, the values of the test genes were predicted from the modes. Figure 3.5 shows the results: for a randomly chosen training set (right), the predictions are quite accurate, whereas a model trained with the first half of samples, corresponding to a subset of the experiments (centre), yielded poor predictions for some genes. Thus, the model does not generalise well between the different experiments, which indicates that some causes of variation behind the data are specific for particular experiments.

For comparison, PCA, ICA, and factor analysis were applied to the cell stress data in a

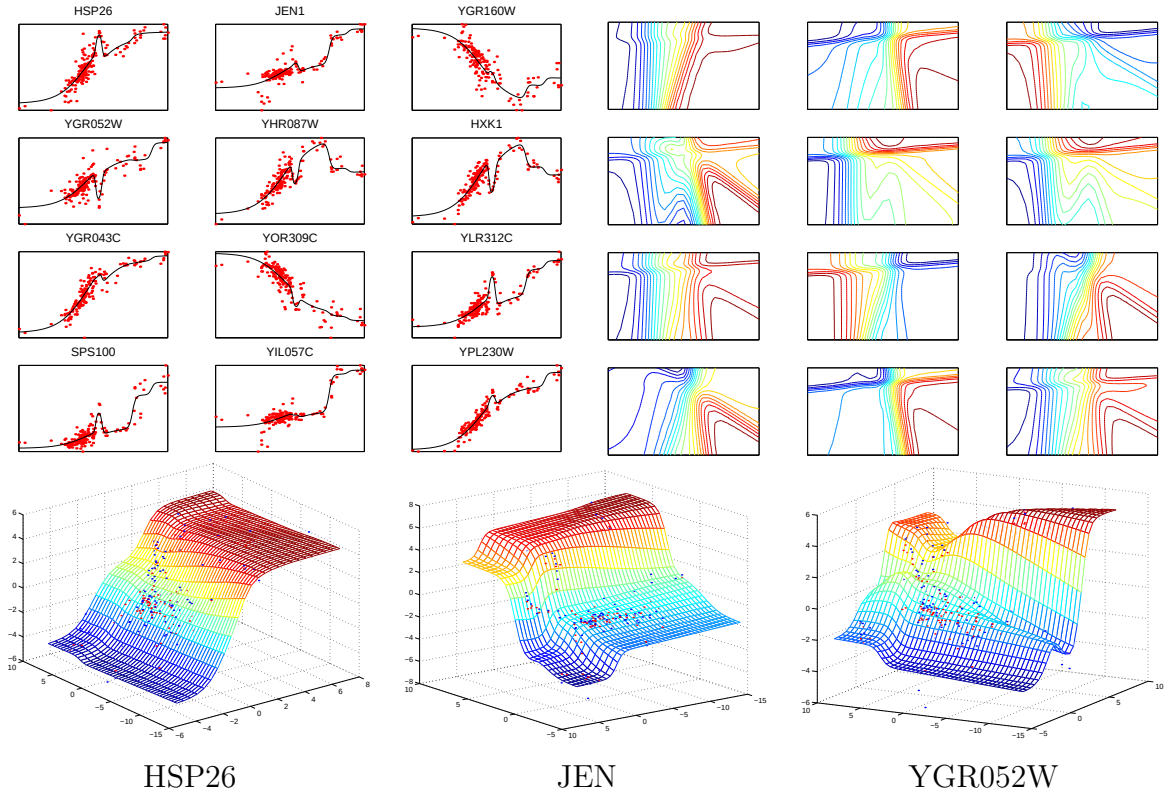


Figure 3.4: Gene programs from the stress response data Gasch et al. [32], determined by nonlinear ICA. Top left: Gene programs for twelve genes. The input variable (on the abscissa) is the single expression mode shown in Figure 3.3, left. The estimated gene program (shown as a curve) interpolates the experimental expression values (dots). Top right: Programs for the same genes, with two nonlinear inputs related to the two modes shown in Figure 3.3, centre. The axes represent the modes, and functions are shown by their contour lines. Bottom: The programs of the genes HSP26 (1), JEN (2), and YGR052W (4) are shown as landscapes. Blue and red dots correspond to experimental and fitted expression values, respectively.

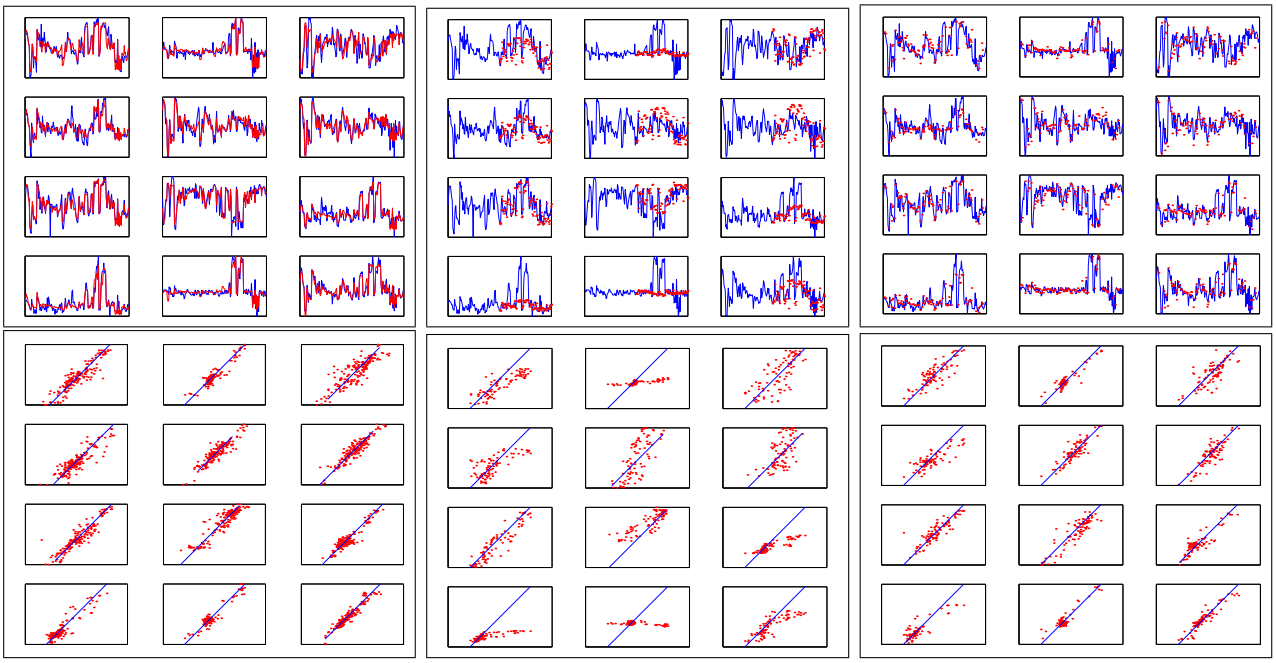


Figure 3.5: Fits and predictions of gene time series from nonlinear ICA. Top left: Time series (blue) of 12 genes (compare Figure 3.4) along with the fits from nonlinear ICA (red) using 2 expression modes. Top centre: The first and the second half of the experimental samples were used as training and test set for cross-validation (see text). Predictions for the test set are shown by red dots. Top right: Randomly chosen training and test sets yield better predictions. Bottom: Same data as above, shown by scatter-plots. The predictions (ordinates) are plotted versus the true values (abscissae). The diagonal lines indicate the true values to be reconstructed.

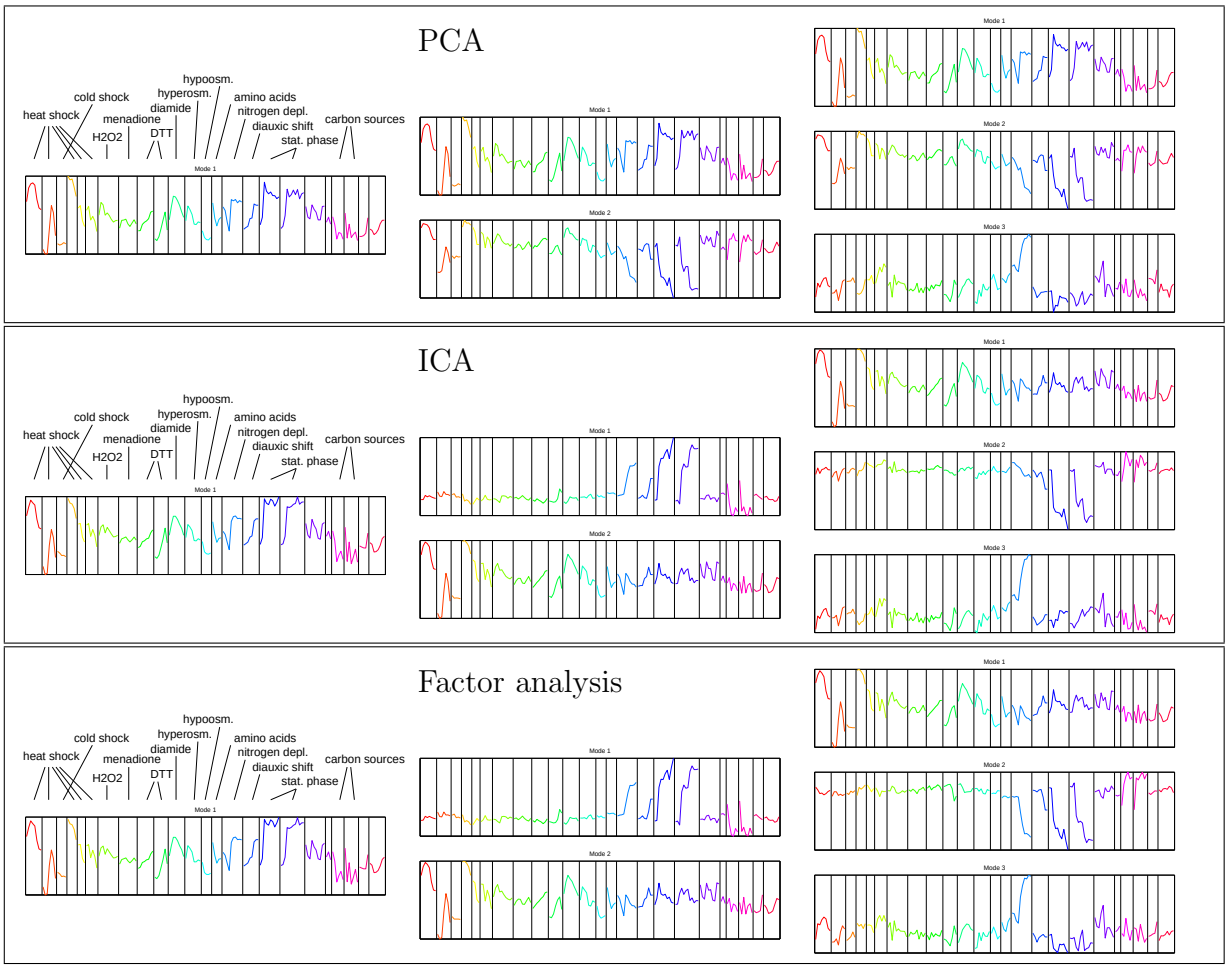


Figure 3.6: PCA, ICA, and factor analysis applied to stress response data Gasch et al. [32]. The diagrams show expression modes (compare Figure 3.3) for nonlinear ICA. Despite their different criteria to separate the modes, the methods yields similar results.

similar manner. These methods are based on different statistical assumptions: with PCA, the linear correlations are supposed to vanish for both influence weights and expression modes, while for ICA, the expression modes are statistically independent, and for factor analysis with the varimax criterion, influence weights are almost sparse. Nevertheless, the modes estimated by different methods resemble each other.

3.3 Biological interpretation of nonlinear ICA

Nonlinear ICA explains expression data by expression modes and nonlinear programs, both estimated from the data. With the expression data studied, the 50 most variant

genes could be explained well by a few modes. Cross-validation shows that the estimated modes do not only fit the data: the model, fitted to a randomly chosen training set, yielded good predictions, so the estimated modes describe systematic, probably biological, effects. However, it is not clear whether the single modes correspond to distinct effects, and how many true biological modes should be expected.

The shapes of the gene programs estimated by nonlinear ICA (Figure 3.4) are more complicated than the ones assumed in the model for simulating expression data. However, in the central region where most of the data points are located, the contour lines are usually almost straight, and in the orthogonal direction, the functions are approximately sigmoidal or curved with a simple shape. The similarity between the modes from nonlinear ICA and different linear methods suggest that nonlinearity is maybe not crucial: in many cases, it mostly accounted for saturation effects at small and large expression values.

There exists a possible application of such estimated gene programs, namely to use them as parts of large cell models. Biologically, expression of genes should be attributed to model variables representing signal molecules, such as transcription factors. However, many regulatory mechanisms are unknown, so possibly, empirical gene programs from a factor model describe expression more accurately than an incomplete ab initio modelling would do. The independence between ICA modes ensures that the single modes contain maximal information. Practically, though, the expression modes would have to be related to the remaining cell variables, as in the cross-validation procedure for nonlinear ICA, in which the modes were calculated from the expression of the training genes.

Chapter 4

Analysis of global gene expression

In this chapter, expression modes are extracted from microarray data by linear factor models. The statistical components are identified with the linear coefficients of the gene programs, rather than with the values of the expression modes. Different linear factor models are tested on simulated expression data. Independent component analysis, which performs well on the artificial expression data, is applied to data from cell cycle experiments and from lymphoma cells. The results of different factor models are compared, and it is studied whether the components describe biological effects.

4.1 Linear components behind global expression

In this chapter, genome-wide expression data matrices X will be analysed by linear models. According to the shape of the data matrix - which usually contains much more genes than experimental samples - statistical assumptions are made about the gene programs, rather than about the expression modes. Thus in contrast to the last chapter, the influence weights of the gene programs are supposed to be drawn from a distribution: this is a convenient way to include prior assumptions about the programs into the model, for instance, the fact that the influence weights should be sparse. Linearity of the gene programs can be assumed if the variation among the experiments is not too large.

Different linear models have been applied to gene expression data, for instance, singular value decomposition [4] [47] which is equivalent to principal component analysis, plaid models [74], correspondence analysis [28], independent component analysis (ICA) [84] [49] [78], and Bayesian decomposition [85]. As an example, Figure 4.1 shows the application of independent component analysis to a gene expression matrix X [107]. The gene profiles (rows of X) can be regarded as points in a multidimensional space with dimensions corresponding to the different samples. The linear decomposition $X = WY$ represents the

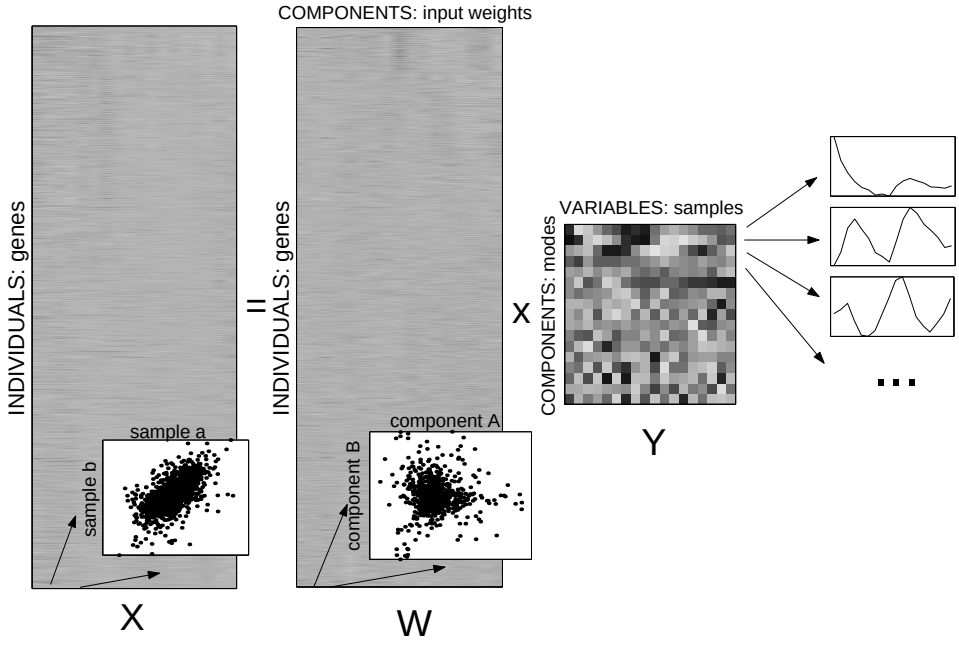


Figure 4.1: Independent component analysis. ICA splits the gene expression matrix X (coded by shades of grey) into a matrix product $X = W\ Y$, introducing new variables (called "independent components" and contained in the columns of W) with minimal statistical dependencies between them. The two lower diagrams show scatter plots between two variables (columns of X) and between two independent components (columns of W). The independent components represent the data with respect to a new basis formed by the rows of the mixing matrix Y. The first three basis profiles are shown in the diagrams on the right. In the context of gene expression, the elements of W and Y may be interpreted as input weights of gene programs and as expression modes, respectively. The data represent a yeast cell cycle experiment [107] where cells show synchronous cell cycle oscillations after treatment with the mating α factor.

data by components (the columns of W, called S in section 1.3) or, alternatively stated, with respect to a new set of basis vectors (the rows of Y, called "mixing" or "loadings" matrix A in section 1.3). The k^{th} expression mode is characterised by its values in the samples (k^{th} row of Y) and by its linear influences on the genes (k^{th} column of W). If logarithmic data are used, the linear combination of inputs corresponds to a multiplicative rather than to an additive processing of signals. Thus in contrast to clustering, linear models account for the combinatorial aspect of gene expression because a gene responds to all modes, in an individual way. The first expression modes (rows of Y) for the cell cycle data in Figure 4.1 show simple, smooth time series. For each mode, groups of genes with high influence weights (in the column of W) can be determined: they are coregulated with respect to this mode, and in analogy to differential expression among samples I shall call them "differentially expressed" with respect to the expression mode.

The components of factor models are separated according to simple statistical assumptions: in the previous chapter, the genes were seen as the variables, like in a mathematical cell model, and it was assumed that the expression modes were statistically independent. Is this assumption plausible? If the modes represent biological signals, then there is at least the argument from information theory: independent signals have optimal coding properties, that is, they can carry a maximal amount of information per time. However, dependence between the expression modes can only be defined for an ensemble of cell states. In data sets containing a limited number of experimental samples, apparent dependencies can be induced by the choice of experiments: in particular, using very similar samples could create dependencies even between biologically independent variables. Factor analysis, on the other hand, assumed sparse loadings, which implied sparse influence weights of the gene programs.

In this chapter, the components are supposed to describe the input weights w_{ik} of the gene programs. PCA (and also singular value decomposition) constrains the modes, as well as the gene input weights, to be orthogonal, i.e., linearly uncorrelated. PCA can be expected to separate a subspace of strong (possibly biological) effects from a subspace of weak noise components, but the biological interpretation of single principal components is not obvious. ICA assumes that different modes exert independent influences on the genes. As a consequence, ICA is sensitive to almost sparse components. The corresponding modes may describe regulators which specifically act on some genes and have little effect on the others. Also nonnegative matrix factorisation yields sparse or almost sparse representations. I also estimated an overcomplete representation using the Laplacian (two-sided exponential) distribution from expression data (not shown), but the results were hardly reproducible. This may be due to the high noise level: also by ICA, only a limited number of components could be estimated robustly. With topographic ICA, the statistical assumption about the gene programs can be stated as follows: if a gene is strongly influenced by a mode, it is also likely to respond to the neighbouring modes. If

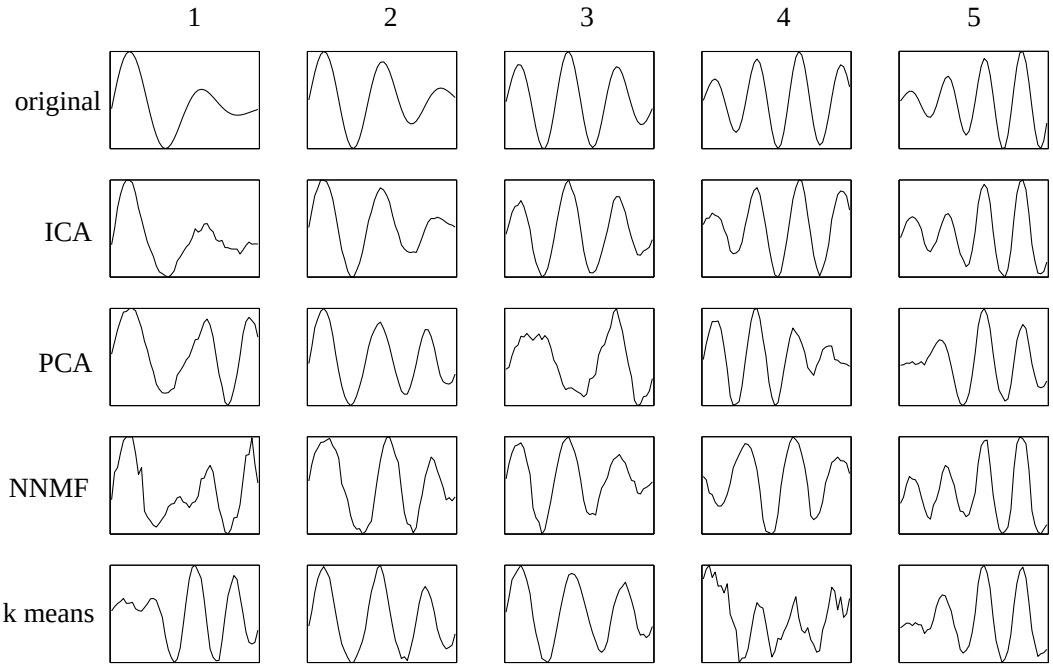


Figure 4.2: Estimating the modes behind artificial gene expression data. Noisy expression data were simulated using the logistic model (section 2.1.2, parameters see appendix, Table B.1 (4)) with five expression modes, which were then reconstructed blindly from the data by linear factor models. Top row: Time series of the true expression modes that were used to simulate the data. The other rows show reconstructions of the time series by independent component analysis (ICA), principal component analysis (PCA), nonnegative matrix factorisation (NNMF), as well as the cluster centres from k-means clustering. For each method, the estimated modes were sorted and scaled such as to match the true modes.

the modes are related to signalling chains, the overlap may indicate cross-talk between them. On the other hand, if the modes are related to cell functions (as it will be proposed in chapter 5), such overlaps may indicate relations between functional systems of the cell, such as different parts of metabolism. Finally, also nonlinear ICA may also be applied to a whole matrix of microarray data, treating the samples as the statistical variables, but the nonlinear functions will not have any obvious interpretation then: the model would assume different nonlinear functions for each sample, but the same for all genes.

4.2 Reconstructing the modes behind simulated data

Can factor models reconstruct expression modes blindly from data? Simulated expression data (see section 2.1.2) with five true modes were used to test this. As the model is

only weakly nonlinear, the modes reconstructed by linear methods can be compared to the original modes. Expression data were simulated for 50 experimental samples and 500 genes, 300 of which respond to the expression modes, with 3 inputs per gene on average. Further parameters are listed in Appendix B.1, (4). The time series of the 5 underlying components are shown in the top row of Figure 4.2. Four linear factor models were tested, namely independent component analysis (ICA), principal component analysis (PCA), nonnegative matrix factorisation (NNMF), and a linear model based on k-means clustering. NNMF was applied to original (not the log-tranformed data), and hypothetical components were determined by a linear regression between the centred data and these logarithmic modes. For k-means clustering, components were determined similarly, regarding the cluster centres as expression modes. The modes from each method were sorted and scaled to match the true modes as close as possible. Figure 4.2 shows that the original modes could be reconstructed more or less by all methods. The results of ICA and PCA were quite reproducible for repeated runs, while NNMF and k-means yielded varying results (not shown).

What happens if too few or too many components are to be estimated? It might be expected that the true modes become mixed or split into several modes, and the estimated modes deviate from the original ones. As a test, the calculations described above were repeated for 5 true components and varying numbers $n = 2, .., 10$ of estimated components. Figure 4.3 shows how the modes estimated by ICA, PCA, NNMF, and k-means match the true modes. The similarity between a true mode and the best matching estimated mode is quantified by the linear correlations between them (left). On the right, correlations between the corresponding components are shown. From the four methods, ICA reconstructed the modes most reliably.

Finally, it was tested how the results of ICA vary with nonlinearity and noise in the data. Artificial data were produced using five different parameter sets. The standard parameters are given in the appendix, Table B.2, (4). The Figures 4.4 and C.1 show correlations between the modes and the components, respectively. Each box corresponds to a different parameter set (see Table B.2 in the appendix): the left and the right column of parameters correspond to a weak or a strong effect of the nonlinearity. The boxes in each column show the results of different parameters, namely a standard parameter setting, a setting without noise, with strong noise, with values in the upper saturation regime, and in the lower saturation regime. Altogether, ICA performs quite well even for noisy data and for wrong numbers of components unless too many expression values are in the saturation regime (see Figure C.1).

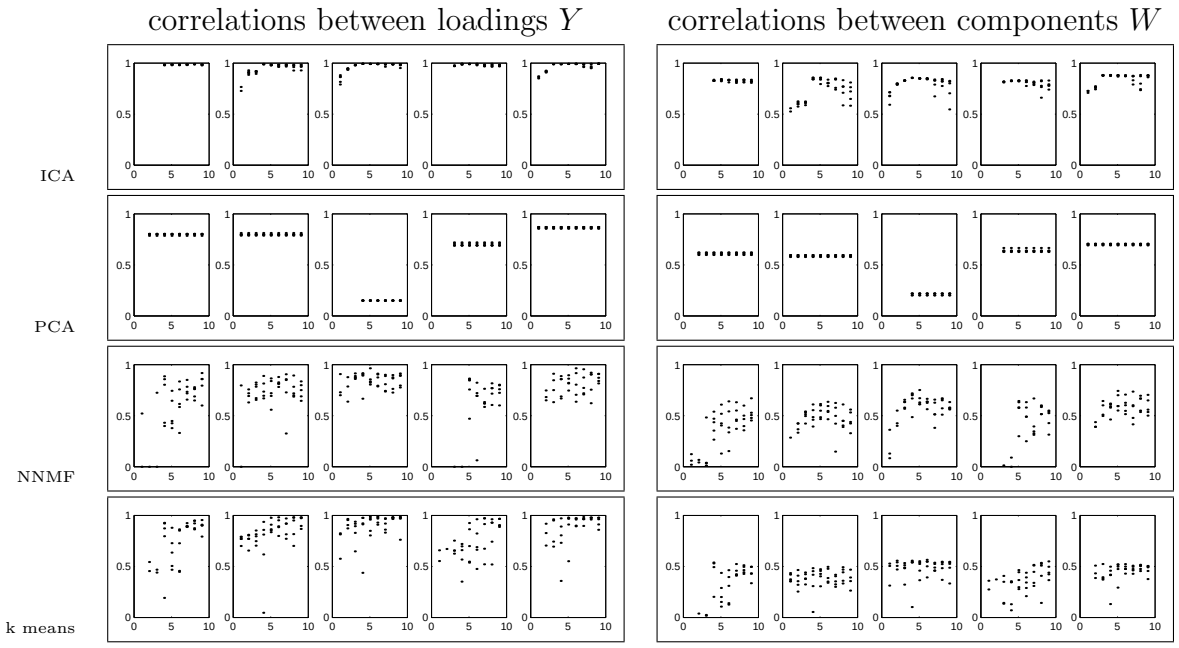


Figure 4.3: Linear models with different numbers of estimated components. The same factor models as in Figure 4.2 were applied to simulated data. For five true components, $n = 2, .., 10$ components were estimated by each method. The similarity between true and estimated components was quantified by the linear correlation between them. Left boxes: Correlations between the true modes and the expression modes estimated by ICA, PCA, NNMF, and k-means clustering. Each small box refers to one of the true modes, and the abscissa denotes the numbers $n = 2..10$ of modes estimated. For each n, the correlation coefficient with the optimally matching mode is shown on the ordinate: each simulation run was repeated 5 times with a different noise term. Right boxes: Based on the same data, the diagrams show the correlation between the respective true and estimated components describing the gene input weights. In most cases, ICA yields the best reconstruction of both the modes and the components.

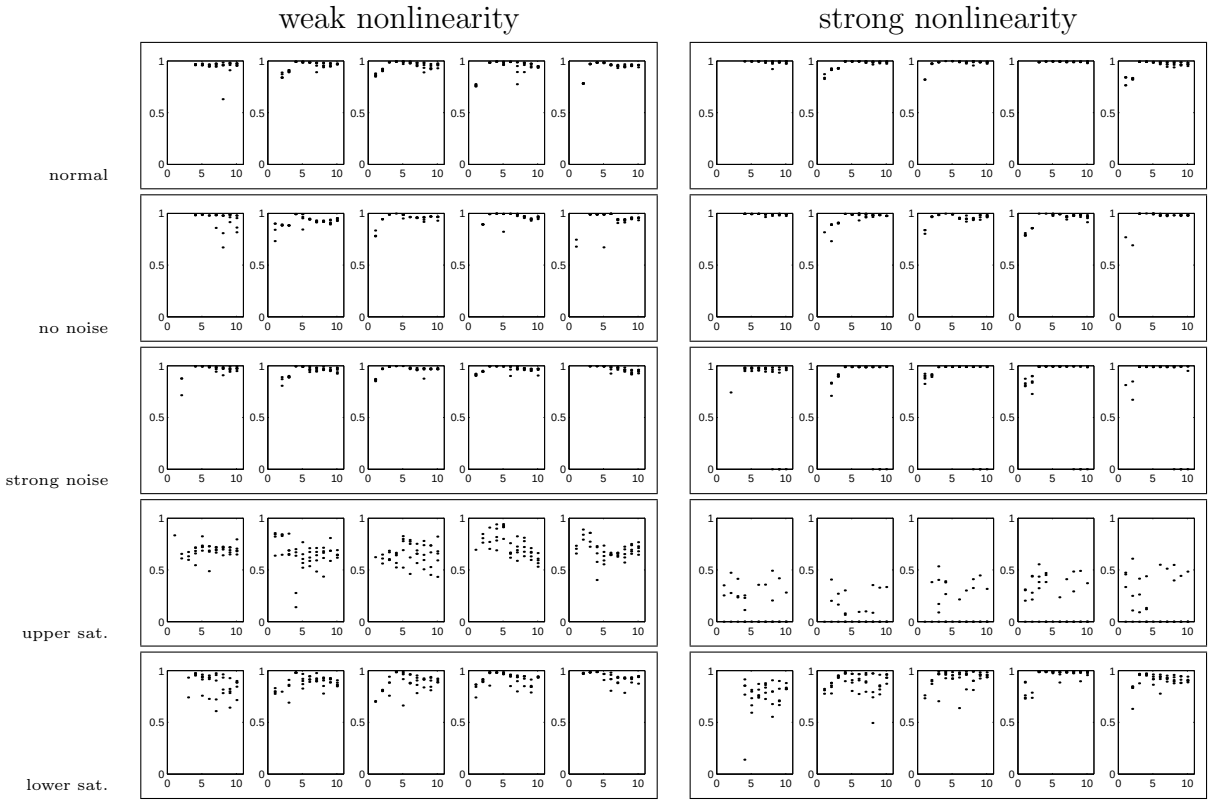


Figure 4.4: Reconstruction of expression modes by ICA, for different parameter choices in the artificial data model. Like in Figure 4.3, $n = 2..10$ components were estimated, and the diagrams show the linear correlations between the five true and the respective estimated modes. Again, each run was repeated 5 times, for different simulations of the data noise. The diagrams refer to different parameter sets (see Table B.2) for the simulation. Left column: Correlations for small variation of the input weights, where the program is almost linear. Right column: Same, with a stronger nonlinearity in the simulated data.

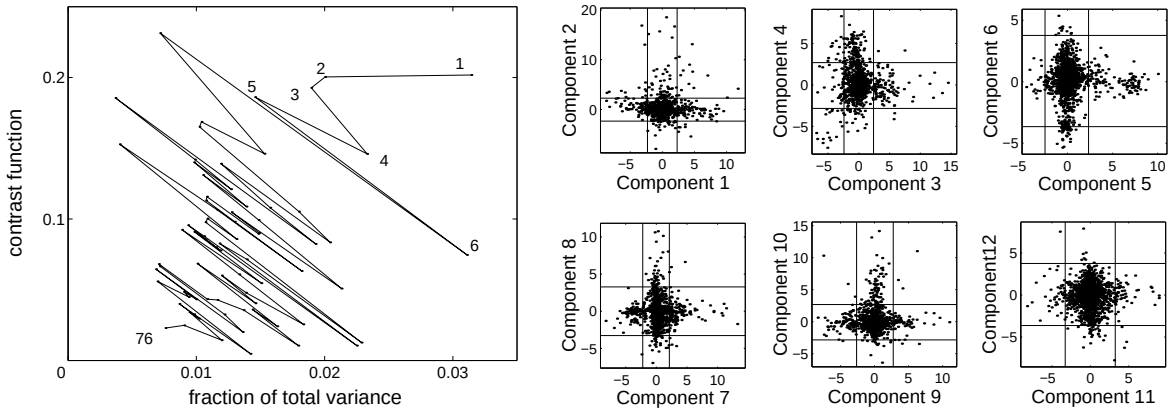


Figure 4.5: ICA of cell cycle data [107]. Left: Sorting the 76 independent components. Each component is characterised by the fraction of the data variance it captures (abscissa) and by a contrast function J_G (ordinate) measuring the non-normality of its distribution. The contrast J_G indicates, among other things, the occurrence of outliers. In the diagram, the components are connected by lines to indicate their order according to a linear combination of both quantities. Components with small values on both axes are likely to represent noise. Right: The first 12 out of 76 independent components. Each box shows the values of two subsequent components plotted against each other, with the gene profiles represented by dots. In these nonorthogonal projections, non-Gaussian structures of the data cloud become visible. For each component, outliers from the normal distribution (thresholds shown as lines) are regarded as strongly induced or repressed genes.

4.3 Analysis of cell cycle experiments

4.3.1 Independent components behind cell cycle data

Spellman et al. [107] studied the expression of 6178 open reading frames (ORF) during the cell replication cycle in the budding yeast *Saccharomyces cerevisiae*. Within separate experiments, cell cultures were synchronised with different methods: addition of the α mating pheromone, which arrests cells in G1 phase, blocking of the cell cycle regulators Cdc15 and Cdc28 [12], and selection of small G1 cells. Moreover, the effects of two cyclins were investigated: Cln3 induces the "start" transition from G1 phase to S phase, when budding and DNA synthesis take place, and Clb2 induces progress through mitosis (M phase), involving separation of the chromosomes and cell division.

The data set contains 77 samples, but shifting the gene mean values to zero confines the data to a 76-dimensional subspace. `FastICA` was applied to the data set. The independent components were sorted as described in Appendix B.3 (also see Figure 4.5, left diagram)

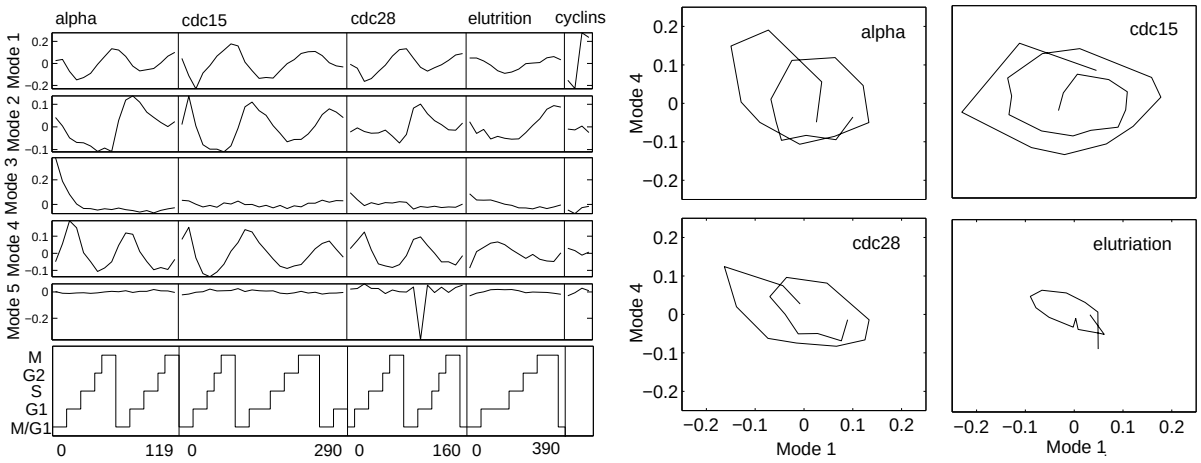


Figure 4.6: Five expression modes calculated from the cell cycle experiments. Left: The samples shown on the abscissa represent time series after cell synchronisation by different methods (mating α factor, *cdc15*, *cdc28*, sorting by elutriation), as well as the activation of cyclins Cln3 and Clb2 (two samples each). The numbers indicate durations in minutes. The cell cycle phases, determined from the cell morphology, are shown in the lower diagram. The corresponding sets of target genes (see Figure 4.5) show that modes 1, 2, and 4 are related to the cell cycle, while mode 3 corresponds to the mating response and according to its influence on the genes (see Table 4.1), mode 5 is related to protein translation. Right: Expression modes 1 and 4, plotted against each other. The four experiments are shown in separate diagrams. Samples are joined by lines to indicate their time order.

with $c = 0.5$, to put similar weight on variance and contrast. The first 12 out of 76 components are shown as scatter-plots in the right diagram of Figure 4.5.

For each component k, sets of induced and repressed genes were determined by the following iterative procedure: the gene with the largest absolute influence value $\max_i(|W_{ik}|)$ was regarded as an strongly responding and excluded until all remaining values were situated within n_σ standard deviations from their median. Thus each mode defined two groups of genes that show a strong positive or negative response. With $n_\sigma = 4$, 2546 genes were found to be strongly influenced by some mode, while about 40 (false) positives would be expected from normally distributed data. Due to their high contrast J_G, the first modes define large sets of target genes, which often contain subgroups related to particular biological functions, mostly consistent with the mode's profile over the samples (see Table 4.1 for a selection of modes). The genes corresponding to lower-scoring modes generally did not share any obvious biological roles.

Cell-cycle behaviour is mainly manifested by the modes 1, 2, and 4, which show a periodic

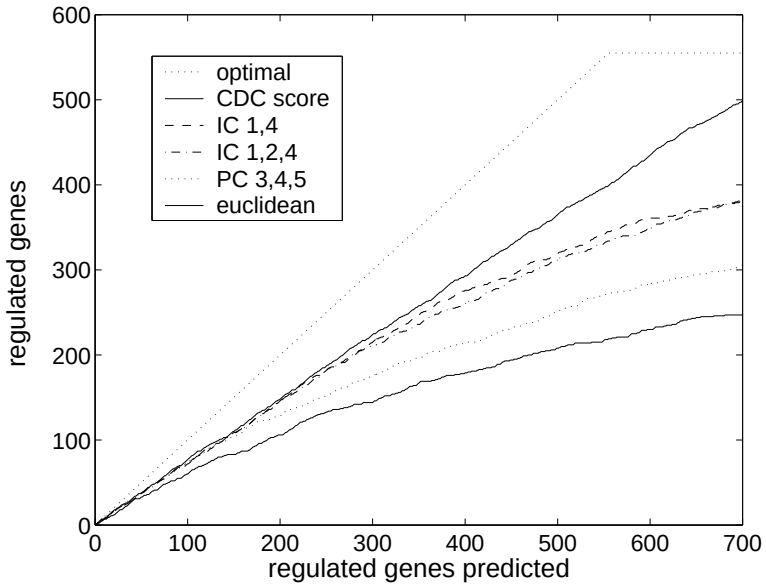


Figure 4.7: Filtering the cell cycle data by ICA improves a prediction of cell-cycle regulated genes. Genes were scored by the variance of their expression the cell cycle experiments, in order to predict 551 genes controlled by known cell-cycle promoter elements. Lower solid curve: the numbers of successful predictions are plotted versus the number of genes predicted. Projecting the gene profiles to the independent components 1, 2, and 4 (dashed-dotted curve) or 1 and 4 (dashed curve) improved the prediction. Projecting the data to the most cell-cycle related principal components 3, 4, and 5 (lower dotted curve) performed less well. The "aggregate cdc score" (see [107]), which compares the expression profiles to sine waves and to profiles of known cell-cycle-dependent genes, yields the best results (upper solid curve). The upper dotted line refers to a perfect prediction.

behaviour with a slow decrease in amplitude, possibly due to desynchronisation (see Figure 4.6). Mode 1, which oscillates between M and S phase, is induced by Clb2 and repressed by Cln3, while mode 4 peaks in early G1 and does not respond in the cyclin experiments. Mode 2 is also active in G1, but remains weak during the first cell cycle round in the α and *cdc28* experiment, and it appears shifted to M-phase during the elutriation time series. In contrast to mode 4, it has a larger influence on metabolic genes than on cell-cycle processes. Mode 3, which reflects the response to the mating α factor, decreases during the G1 phase. Many modes are activated specifically in some of the experiments, or even in single samples. For instance, mode 5 seems to represent an induced protein production in a particular *cdc28* sample and might be filtered out as an experimental artefact.

Alter et al. [4] analysed the same data set by applying singular value decomposition to the separate experiments. To compare the results of PCA and ICA, I generated PCA modes

	Description of the mode	induced functions	repressed functions
1	mitosis vs. replication •	M cyclins, mitosis, MCM complex, cytoskeleton, cell wall, stress, mating cascade, H^+-transport, galactose, secreted acid phosphatases	S phase cyclins, DNA replication, histones, spindle pole duplication, bud emergence, cell wall
2	G1 •	G1/S cyclins, stress, mating, cell wall, lipid production	energy and amino acid metabolism
3	mating response	mating, cell wall, metabolism	G2/M and S cyclins, histones, stress, metabolism
4	replication/budding vs. separation •	G1/S cyclins, MCM, DNA replication/repair, chromatin, subtelomerically encoded genes	G2/M cyclins, histones, cell wall
5	translation	ribosomal, proteins, sugar metabolism	ribosomal
6	growth	cell wall, sugar	RNA processing
7	sporulation •	sporulation, proteins, metabolism	meiosis-specific
10	(single *cdc15* sample)	meiosis, proteins	
11	(decrease in elutr. exper.)	stress, metabolism, Cu/Fe transport	cyclins
14	galactose •	galactose metabolism	hexose transport, sugar
15	(• during *cdc15, cdc28*)	galactose, protein targeting	stress
16	(rising during *cdc15*)	mating α type, stress	
18	(single *cdc15* sample)	meiosis, proteins	
19	late mating response	mating, meiosis, proteins, metabolism	
21	ribosomes	ribosomes	
22	oxidative/osmotic stress		oxidative/osmotic stress, sugar
23	ribosomes (falling in elu.)		ribosomes, translation
24	stress		stress
25	methionine •	methionine metabolism	sugar

Table 4.1: Expression modes from the cell cycle data. For each mode, up- and downregulated genes were selected by thresholding (compare Figure 4.5, right). For many of the modes, a large fraction of these genes is related to particular biological functions which are listed in this table. Modes showing cell-cycle oscillations are marked by a dot.

from the whole data set. With both methods, most of the cell-cycle behaviour is captured by a small number of modes, but the separation into oscillatory, spiky, and noise-like patterns is more distinct with ICA. Besides, the first PCA modes vary within all time series, while various ICA modes remain almost inactive within some of the experiments, although this is not forced by the method.

4.3.2 Dimension-reduction and bootstrapping

Dimension reduction can be used to compress data sets before further calculation-intensive study. Assuming that cell cycle behaviour is sufficiently captured by the modes 1, 2, and 4, one may omit the remaining modes, thereby compressing the data from 76 to 3 dimensions. Moreover, a projection to biologically relevant directions should improve predictions of cell-cycle regulated genes from the expression data. This was tested using a list of 551 genes which are known to be controlled by cell-cycle promoter elements (taken from the web supplement of [107]). All ORF were scored by the variance of their expression levels in the cell-cycle experiments, and the n_{pred} highest-scoring genes were predicted to be contained in the list. Figure 4.7 shows the number of successful predictions as a function of n_{pred}, based on the original data as well as on different kinds of filtered data: projecting the profiles to the cell-cycle-related principal components 3, 4, and 5 improved

the prediction considerably, but replacing the gene expression profiles by the cell-cycle related independent components yielded an even better prediction. The best prediction was achieved using the "aggregate cdc score" [107], which compares the gene expression profiles to sine and cosine waves and to the profiles of known cell-cycle regulated genes.

Does `FastICA` yield robust results for different seeds of the random number generator and for different choices of the genes and the experimental samples? Figure 4.8, top left, shows how the expression modes varied among estimation runs for the full data set. In contrast to above, only twelve modes were estimated. The modes from different estimation runs were sorted and scaled to be comparable to each other. For each mode, the mean and the standard deviations over ten estimation runs are shown in the diagram. The other boxes show analogous results for different runs where genes, experimental samples, or both had been resampled. For bootstrapping the experiments, 30 resampled experiments (out of 77) were used in each run, while for bootstrapping the genes, 3000 genes were resampled out of 6178. For most of the modes, the estimated errors are rather small. If both genes and samples are resampled, the modes 1,2,3,4, and 6 remain stable, while the experimental artefact in mode 10 (see below, section 4.3.3) has almost disappeared. The distribution of the components yields also error estimates for the input weights w_{ik}: they can be used for testing which of the input weights differ significantly from zero (not shown).

4.3.3 Comparing different factor models

Figures 4.9 and 4.10 show components estimated by different methods, namely ICA [56], PCA, topographic ICA [56], NNMF [75], k-means clustering, clustering by the linear correlation, and generalised canonical analysis. The components from each method were sorted to match the ICA modes: the criterion was to maximise the sum of squared linear correlations. Except for PCA and canonical analysis, all methods are supposed to be sensitive to almost sparse components. With each method, 12 components were estimated: the 12 components for ICA differ from the first 12 out of 76 components considered in the previous section. Clustering by the linear correlation between the gene profiles was done in a similar manner as k-means clustering: the clusters are represented by prototypes chosen from the gene profiles, such that the mean squared correlation with the members of the cluster is maximised. In addition, ICA was applied to differently preprocessed data: besides the usual log-transformation (Figure 4.9, row A), the logistic transformation was used (row B, for each gene, x_∞ was chosen to be 1.2 times the maximal value), and untransformed data were considered (row C).

Some features of the modes, such as the oscillations in mode 1 and 4, are quite consistent among the methods. In particular, ICA (A) and k-means clustering (G) yielded similar

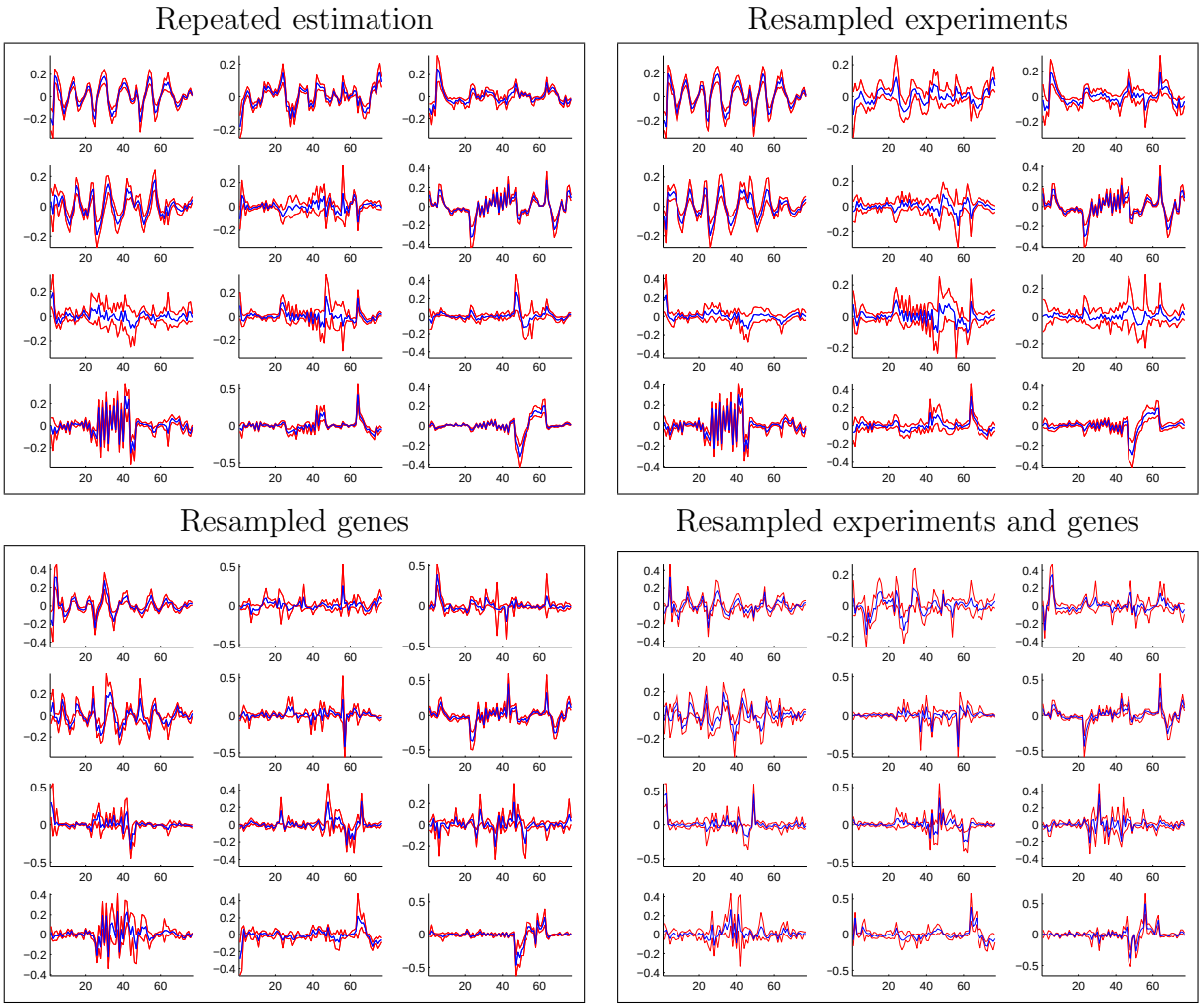


Figure 4.8: Bootstrapping the results of ICA. Twelve independent components were estimated from the cell cycle data. Each diagram shows average expression modes (rows of A) and their standard deviations, determined from 10 estimation runs. Top left: Repeated estimation with different random seeds for the FastICA algorithm. Top right: Bootstrapping the experiments. From the 77 experimental samples, 30 were randomly chosen in each resampling run, to study estimation errors due to the choice of experimental samples. Bottom left: Bootstrapping the genes. Bottom right: Bootstrapping both the samples and the genes.

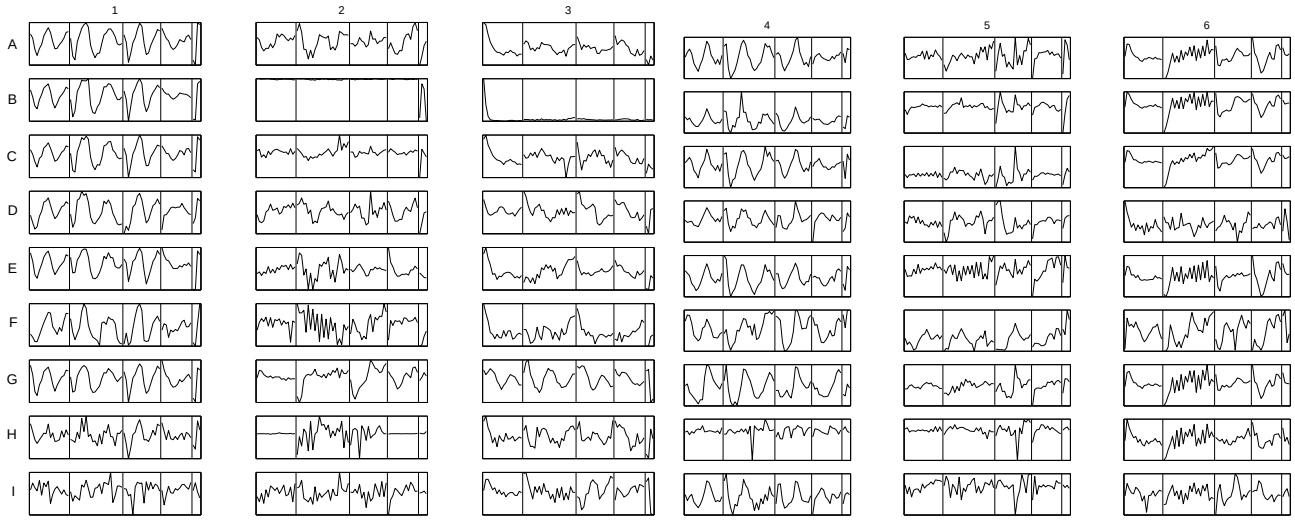


Figure 4.9: Comparing different linear models for the cell cycle data. Each row of diagrams shows time series of the first six expression modes from different methods, namely (A) ICA, (B) ICA without log-transformation, (C) ICA with logistic transformation, (D) PCA, (E) topographic ICA, (F) NNMF, (G) k-means clustering, (H) correlation clustering, and (I) canonical analysis. With each method, the modes have been sorted to match the ICA modes. The modes 7-12 are shown in Figure 4.10.

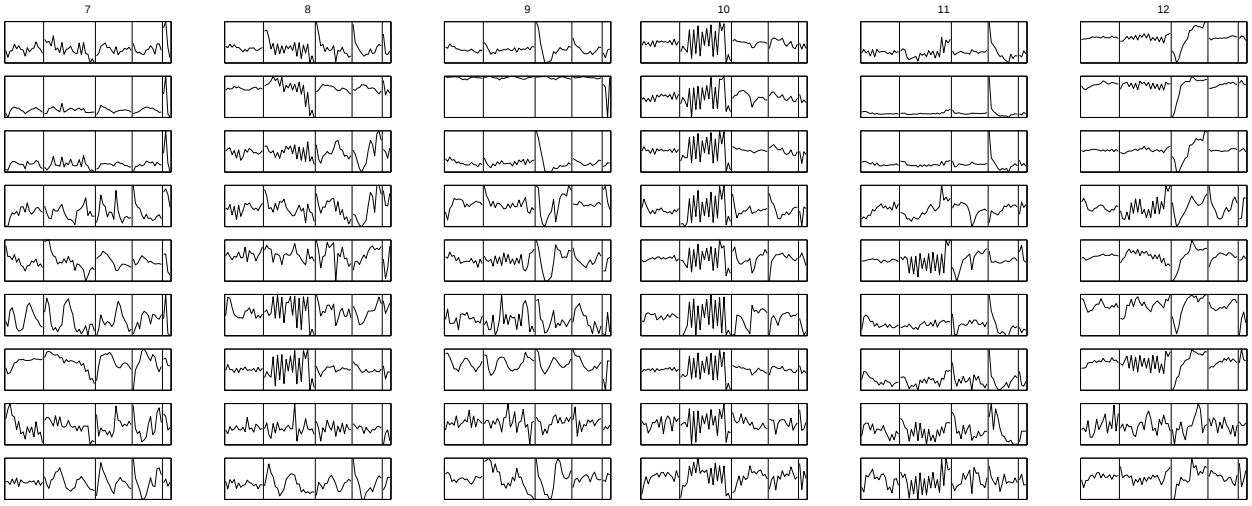


Figure 4.10: Same as Figure 4.9. Here the modes 7-12 are shown.

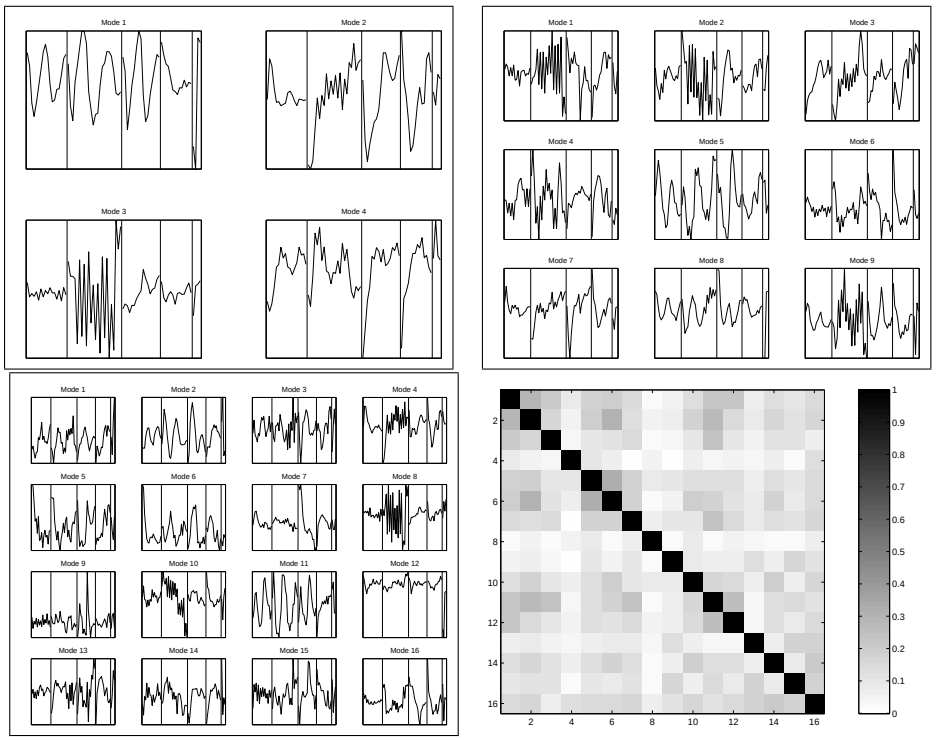


Figure 4.11: Cell cycle modes from topographic ICA. The components (not shown) are arranged in a predefined topology such that the higher-order dependencies are concentrated between neighbouring components. Neighbouring modes tend to influence similar sets of genes, but with uncorrelated influence weights. The boxes show expression modes for quadratic grids of side length 2, 3, and 4. Bottom right: Correlations between the absolute values of topographic independent components from the 4×4 grid. Dark local shapes in the matrix indicate the dependencies between neighbouring components.

results. If the data are preprocessed in different ways (rows A, B, C), most ICA modes are still recognisable: The modes 1, 4, 6, 7, 10, 11, and 12 depend only weakly on the preprocessing, while the others show larger differences. The results of canonical analysis (I) and clustering by linear correlation differ strongly from those of the other methods. The results of topographic ICA (H) are shown again in Figure 4.11. The components (input weights w_{ik} of different modes) are linearly uncorrelated, while their absolute values may be correlated between neighbouring modes. In contrast to usual ICA, the objective is not to minimise the higher-order dependencies, but to concentrate them (in particular, the correlations between the absolute values of components) between neighbouring components. The large boxes show expression modes for different numbers of components arranged on quadratic grids of side length 2, 3, and 4.

Variation in the data is not necessarily of biological origin, but may also be due to experimental artefacts: in the cdc15 experiment, fast periodic oscillations were detected by

almost all methods (mode 10 in Figure 4.9). The explanation for its oscillatory shape is simply that the even- and odd-numbered samples were hybridised on different days. Possibly, some genes responded to the slightly different sample preparation or hybridisation conditions. In the latter case, the genes concerned do probably not share any biological role. This may be the reason why ICA, by its independence assumption, can separate the artefact from other processes.

4.3.4 Expression modes and gene functions

To interpret the ICA modes from the cell cycle data, both the modes themselves and the gene lists from respective components were taken into account: the oscillatory shape of some modes suggested that they might describe cell cycle effects, and this interpretation was supported by the high fraction of cell-cycle-related genes responding to these modes. In the following, such relations between expression components and the gene functions are tested quantitatively.

Some of the ICA modes in Figures 4.9 and 4.10 appear related to particular cell cycle experiments. To test this observation, each mode was characterised by a profile describing their activity in the different experimental time series[1]. Figure 4.12 shows these experimental profiles for six of the methods from Figure 4.9. The significant[2] profiles at the 1% level are indicated by a cross. ICA, topographic ICA, and k means clustering show particularly many significant profiles.

If an expression mode represents biological processes, the corresponding genes should share common functions. This was tested quantitatively for the linear models shown in Figure 4.9: for each mode, strongly responding genes were determined by the thresholding method described in section 4.3.1, with $n_\sigma = 3$. The resulting gene groups were related to 95 MIPS functional categories [82]. The pairs with strong associations[3] ($m > 3$ and $n_{ik} > n_{ik}^{exp}$) are listed in Table 4.2. Eleven pairs (expression mode/MIPS category) were found for ICA, and 17 for k-means clustering, while under the null hypothesis, only 1.7 combinations would be expected on average.

[1] After subtracting the median from the mode's time course in all samples, the sum of squares was calculated for each of the five cell cycle experiments (including the cyclin experiment) and normalised by the total sum of squares.

[2] Significance of each profile was studied by comparing its maximal value to the results of a permutation test (1000 permutation runs).

[3] Genes were counted for each pair (expression mode/MIPS category), and the actual count number n_{ik} for mode i and category k was compared to the number $n_{ik}^{exp} = n_i n_k / n$ expected under the null hypothesis (no dependence). High or low count numbers were detected by the test statistics $m = (n_{ik} - n_{ik}^{exp})/\sqrt{n_{ik}^{exp}}$ which is close to normal for large $n_{ik}^{(exp)}$.

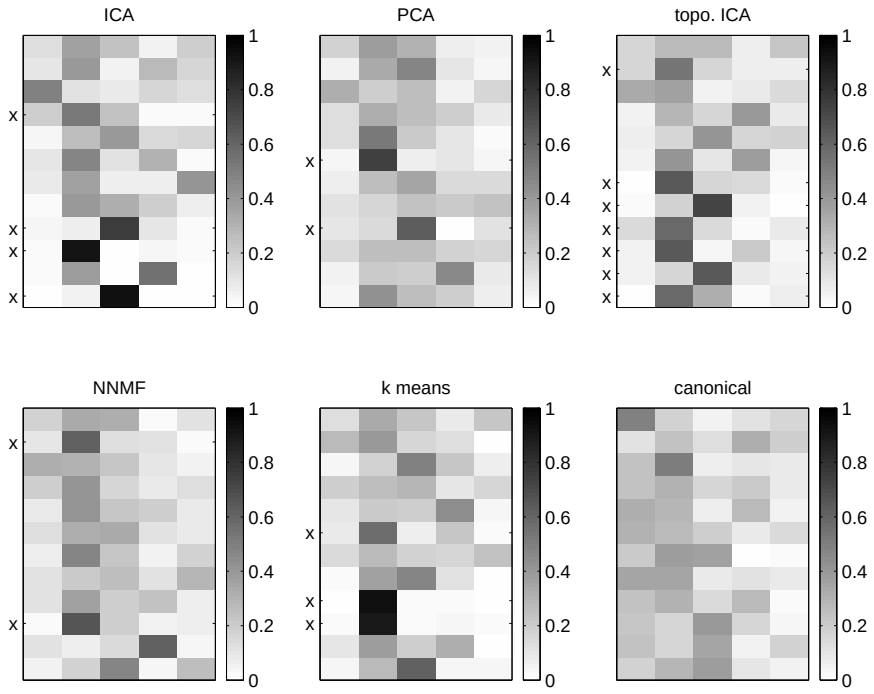


Figure 4.12: Experimental profiles indicate which modes are specific for some of the cell cycle experiments. The matrix rows show experiment profiles for the modes from six linear methods (compare Figures 4.9 and 4.9). The experimental profile indicate whether the squared values of a mode are concentrated within some of the five experimental time series. Crosses indicate significant profiles (at 1% level), where the largest element is higher than expected from a permutation test.

	Mode	m	Genes	Functional category
ICA	5	14	55	CYTOPLASM
	6	7.8	8	RRNA TRANSCRIPTION
	8	4.4	10	MITOCHONDRION
	1	4	18	DNA PROCESSING
	1	3.6	9	CENTROSOME
	4	3.5	45	NUCLEUS
	7	3.4	7	EXTRACELLULAR / SECRETION PROTEINS
	6	3.1	5	OTHER TRANSCRIPTION ACTIVITIES
	9	3.1	9	STRESS RESPONSE
	1	3.1	21	CELL CYCLE
	12	3	88	UNCLASSIFIED PROTEINS
PCA	6	3.3	13	NUCLEUS
TICA	5	4.3	14	CYTOPLASM
	1	3.9	17	DNA PROCESSING
	4	3.2	5	CELL WALL
	4	3.1	13	CYTOPLASM
NNMF	9	4.1	8	MITOCHONDRION
	4	4	8	DNA PROCESSING
k means	6	6.1	14	RRNA TRANSCRIPTION
	4	6	16	DNA PROCESSING
	7	5.8	14	ENDOPLASMIC RETICULUM
	11	5.6	21	MITOCHONDRION
	7	4.7	7	INTRACELLULAR TRANSPORT VESICLES
	7	4.2	6	NITROGEN AND SULFUR METABOLISM
	12	4.1	231	UNCLASSIFIED PROTEINS
	1	4.1	7	PLASMA MEMBRANE
	6	3.5	39	CYTOPLASM
	7	3.4	11	CYTOSKELETON
	7	3.3	6	CENTROSOME
	7	3.3	6	GOLGI
	11	3.3	32	CYTOPLASM
	4	3.2	41	NUCLEUS
	4	3.1	13	CELL CYCLE
	6	3.1	6	RIBOSOME BIOGENESIS
	11	3.1	6	PROTEOLYTIC DEGRADATION
corr. k means	7	6.6	13	EXTRACELLULAR / SECRETION PROTEINS
	5	4.7	44	CYTOPLASM
	3	4.2	6	PEROXISOME
	6	3.7	111	CYTOPLASM
	6	3.6	10	RRNA TRANSCRIPTION
	1	3.5	19	DNA PROCESSING
	7	3.1	9	CENTROSOME
	1	3.1	8	CENTROSOME
	12	3.1	6	ALLANTOIN AND ALLANTOATE TRANSPORTERS

Table 4.2: Correspondence between expression modes and MIPS functional categories [82]. For each mode, strongly responding genes were determined by thresholding (see section 4.3.1). Each set of strongly responding genes was compared to each MIPS category by counting the genes contained in both of them. A statistical variable m (see text) measures the deviation between the count number and the number expected from the null hypothesis (no correspondence between modes and functions). The table shows significant sets (with $m > 3$) of genes that are differentially expressed in a certain mode and belong to a certain functional category. Sets containing fewer than 5 genes have been omitted.

Finally, coregulation within parts of the metabolic network was studied. With ICA, large sets of coregulated genes will give rise to an independent component. If genes on metabolic pathways are coregulated, the largest values of the corresponding component should be concentrated in particular regions of the metabolic network. If a network distance D_{il} is defined between the genes, the localisation of the k^{th} component with values $w_{\cdot k}$ can be measured by the correlation function

$$c^{(i)}(d) = \frac{< w_{ik} w_{lk} >_{(i,l)}}{< w_{ik}^2 >_i} \tag{4.1}$$

where the indices (i,l) run over all pairs of genes in the distance $D_{il} = d$. Figure 4.13 shows the results for the cell cycle data[4]. For most of the components, the correlation significantly decreases with distance: for comparison, the bars show 10% and 90 % quantiles of the correlation function from a permutation test[5]. The coregulation of subsystems becomes better visible in the stress response data Gasch et al. [32] (see Figure 4.14). Twelve ICA modes were estimated, and two of them are localised at glycolysis or the TCA cycle, so the components indicate a coregulation of genes within these subsystems.

4.4 B-cell lymphoma data

ICA was applied to a second data set related to different cell types rather than to time courses. Alizadeh et al. [2] investigated the expression of 4026 human genes in 96 samples of normal and malignant lymphocytes. The "lymphochip" used in this study contains clones from lymphoid cDNA libraries as well as genes related to immune-response and oncogenesis. The samples included T cells, activated blood B cells, B cells from the germinal centre (GC), six leukaemia cell lines (WSU1, Jurkat, U937, OCI Ly12, OCI Ly13.2, SUDHL5), and cells from three types of lymphomas: follicular lymphoma (FL), chronic lymphocytic leukaemia (CLL), and diffuse large B cell lymphoma (DLBCL). The cell samples are shown in Figure 4.15 by scatter-plotting the first 12 out of 95 expression modes (see also Table 4.3). The modes were compared to the gene clusters that had been determined in the original work using hierarchical clustering. Although the clusters and modes are not directly comparable (as the modes describe additive effects), some of them seem related: modes 2 and 5, which show the highest variance among the modes, point towards the "proliferation" and "lymph node" gene clusters , while mode 8 and 12 are related to the "pan B cell" and the "germinal centre B-cell" cluster, respectively. Alizadeh

[4]Yeast ORF and the corresponding chemical reactions were mapped via the EC numbers (see Appendix B.4), yielding hypothetical EC expression data. A network distance between EC numbers was defined by yeast metabolic network described in chapter 8.

[5]In the test, the correlation functions were calculated for expression data with randomised order of the EC numbers.

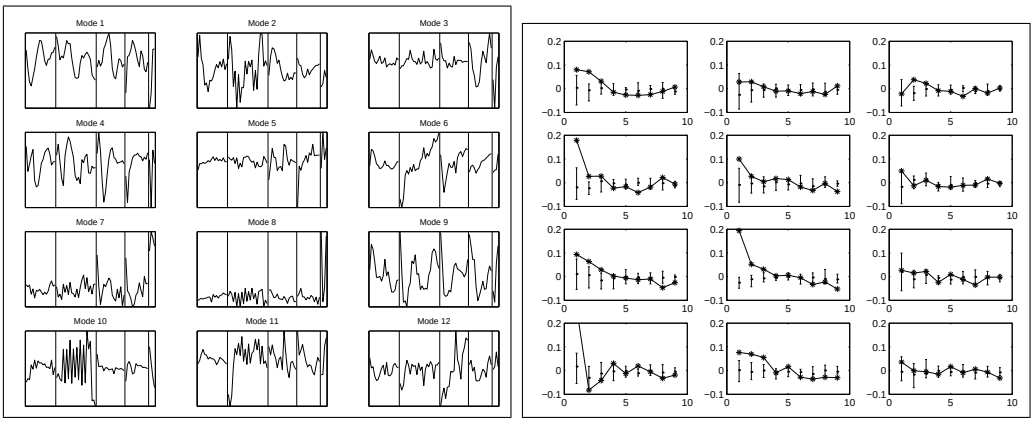


Figure 4.13: Localisation of independent components in the metabolic network. Cell cycle expression data were averaged over genes related to the same EC number. A distance D_{ik} between EC numbers was defined by the minimal network distance of the corresponding chemical reactions (see chapter 8). Left: Twelve ICA modes for the EC expression data. Right: Correlation functions (see text) for the twelve independent components indicate their localisation in the metabolic network. Except for the modes 3 and 9, the correlation functions decrease, so close reactions tend to be coregulated by the expression mode. A permutation test shows that this decrease is significant: the bars represent 10% and 90% quantiles over 15 permutation runs.

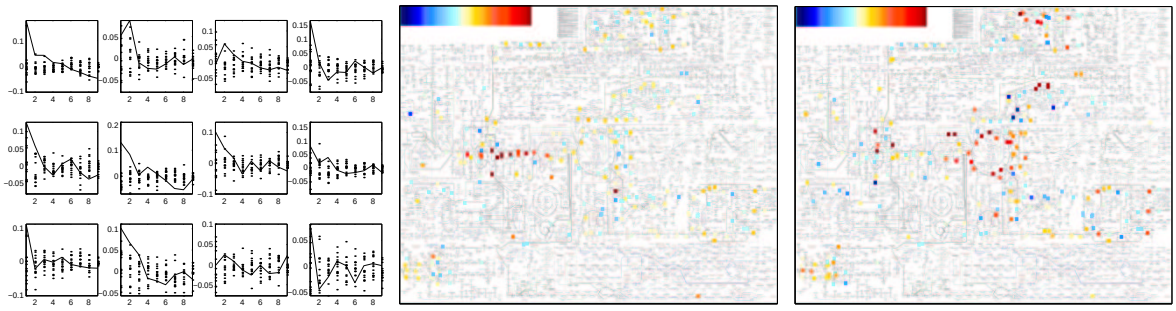


Figure 4.14: Twelve ICA modes from the stress response data Gasch et al. [32]. Left: Correlation functions. Most of the twelve components are significantly concentrated in the metabolic network (compare Figure 4.13). Two of them are shown on the Boehringer chart. They are localised in small regions of metabolism. High absolute values are concentrated in glycolysis (centre, mode 7) and the TCA cycle (right, mode 2).

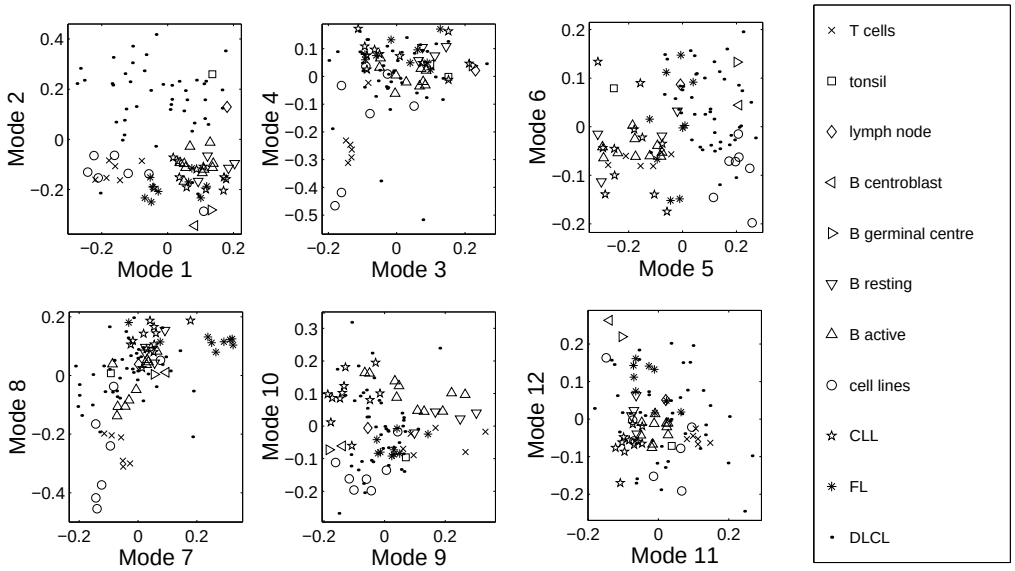


Figure 4.15: Cell samples from the lymphoma data (Alizadeh et al. 2000) [2]. The axes represent levels of the first 12 ICA expression modes, each diagram showing a projection to two subsequent modes (rows of A). In some of the projections, clusters of cell types (indicated by different symbols) become visible. A description of the modes in terms of related cellular functions is given in Table 4.3.

et al. stated that genes from "T cell signature" appearing in DLBCL samples indicated the presence of T cells in the biopsies. Mode 11, which is related to this cluster, may be expected to describe the contamination with T cells and might be filtered out to correct the DLBCL expression patterns for this particular effect.

4.5 Conclusions

In the previous two chapters and in this one, it was shown how gene expression data can be explored by factor models: with the linear methods, the data were represented by linear superpositions of effects, each characterised by a component and their respective loadings. With nonlinear ICA, a nonlinear mixing of the components was realised by a neural network. As in previous expression analyses by other authors, the objective was to determine coregulated genes and to study whether coregulation of genes implies common biological functions. Clusterings and linear models implement different concepts

	Mode	upregulated in	downregulated in	functions induced	fct. repressed
1	B cell activation	lymph node, tonsil, blood B, CLL, SUDHL6	T cells, Jurkat, U937, OCI	immunoglobulins, differentiation	
2	lymph node	DLBCL, lymph node, tonsil		interferon-induced genes, activation, defence	
3		lymph node, tonsil, germinal centre	T cells, Jurkat, U937, OCI	immunoglobulins	
4	MHC		T cells, Jurkat, U937	MHC	
5	proliferation	DLCL, cell lines, germinal centre	T cells, active B CLL, tonsil	cell cycle	interferon-inducible
6		DLCL, germinal centre tonsil, lymph node		immunoglobulins	
7	FL	FL		anti-proliferative	
8	B vs. T cells	CLL, FL	Jurkat, OCI, T cells	B receptors	T receptors
9		blood B, T cells SUDHL6, Jurkat, U937	germinal centre, CLL	adhesion, proliferation, shock, signalling	
10		blood B, CLL	cell lines	B receptors	
11	T activation	T cells	active B, FL, CLL	T activation, chemokines, T receptors (CD3), interferon-inducible genes	adhesion
12	germinal centre	germinal centre, FL	OCI	B activation	homing

Table 4.3: The first 12 expression modes inferred from the lymphoma data. Modes are characterised by the cell types in which they are most up-/downregulated (compare Figure 4.15) and by functions of their target genes.

of coregulation: distance measures between genes, like the correlation or Euclidean distance used in clusterings, compare the gene profiles as a whole, that is, two genes are considered coregulated if they behave similarly in all experiments. This assumption is partially relaxed in biclusterings, where genes only need to show similar profiles in some of the experiments. The linear factor models assume unobserved expression modes that influence the expression of many genes and that are superposed to yield the gene profiles, in analogy to signals that are integrated to regulate the transcription of genes. In reality, gene expression is controlled by a combination of biochemical signals, such as regulatory proteins. For the purpose of modelling, though, it can be attributed to general variables ("expression modes") characterising the cell state. This schematic view of gene expression was used here for the simulation of expression data, and statistical models of the same mathematical form were fitted to experimental expression data.

The real processes behind expression data are certainly too complicated to be reconstructed from noisy and incomplete data. In a series of experiments, those modes which vary considerably among the samples will appear in the data and may be reconstructed by a factor model, while the others are buried in the measurement noise. Can we expect that true biological modes can be identified by blind estimation? Even with appropriate criteria to separate the components, there remain some restrictions: if the number of true modes exceeds the number of samples, they cannot all be resolved, and if the values of two modes, or the corresponding input weights, are statistically dependent, it can be difficult to separate them. Therefore it is important to use data from diverse experimental conditions: only if the modes vary independently among the samples, are they likely to appear as separate factors. It is not necessary, though, that the individual modes are

activated specifically only in particular experiments. Once components behind the data have been determined, they can be used for several purposes: the individual components and expression modes can be visualised as if they were profiles of hypothetical samples or genes, and the modes can be used as features for the discrimination of samples [84]. Projecting the data to some of the components leads to a new, problem-specific data metric that can highlight aspects of special interest, while information about all genes and samples is maintained.

Chapter 2 contains preliminary tests in which linear and nonlinear models were fitted to simulated expression data, based on simple sigmoidal gene programs. Linear models failed to handle the saturation for high or low values, but the nonlinearity could be reduced by an appropriate preprocessing of the data. The regression models were robust against additive and multiplicative noise in the data. Nonlinear gene programs and the corresponding expression modes were estimated blindly in chapter 3 by nonlinear ICA. A variant of cross-validation was used to test the results against over-fitting: cross-validation between different experiments yielded worse predictions than for randomised experimental samples, which indicates that processes described by the modes are specific for some of the experiments. For the cell cycle data, this hypothesis was confirmed in section 4.3.4 by the experimental profiles.

To be useful for expression analysis, factor models are supposed to separate biologically significant components from components that represent noise, experimental artefacts, and irrelevant biological processes. The factor models considered here separate their components according to predefined statistical criteria. Nonlinear ICA in chapter 3 assumed statistical independence among the expression modes as a criterion to separate them. However, dependence among the modes can only be defined for an ensemble of cell states and depends on the choice of experimental samples studied. Thus in chapter 4, the samples rather than the genes were regarded as the statistical variables. Accordingly, the components represented the linear input weights of gene programs, while their loadings corresponded to the profiles of expression modes. Independence between the components, which is assumed by ICA, lead to biologically sensible results. When blind linear methods were tested with simulated expression data, ICA detected the underlying modes quite reliably. Real biological modes, though, might be much sparser and more correlated than assumed in the simulation, and if the noise level is high, weakly varying biological modes may not be identified. Applied to the cell cycle data, ICA separated oscillatory modes from other effects, like the decreasing influence of the α-factor used for synchronising the cells. The oscillatory modes were interpreted by cell-cycle processes, and this was also confirmed by the annotations of the respective genes and by the fact that projecting the data to these modes improved the prediction of cell-cycle-related genes. Microarray data contain experimental artefacts, which can be detected by factor models: for instance, in the lymphoma data set, an expression mode could be attributed to the contamination

with T-cells, while mode 10 from the cell-cycle data probably represents a hybridisation effect. The relations between statistical components and gene functions were also shown quantitatively: projecting the data to the oscillatory ICA modes improved a prediction of cell-cycle-related genes. Strongly influenced genes of individual ICA modes share functional annotations and appear concentrated in the metabolic network, both with the cell cycle data and with the cell stress data [32] shown in Appendix C.

Part II

Optimal differential expression

Chapter 5

Analysis of optimal differential expression

This chapter is concerned with the optimal behaviour of biological regulators, in particular differential expression of genes and regulation of enzyme activities. A mathematical model is proposed in which the regulators control steady-state properties of the cell. Their behaviour is governed by an optimality principle: they adapt themselves to any external condition in a such way that an objective function is maximised. According to the model, the differential expression of a gene after a small perturbation reflects two quantities, namely its effect on important cell variables and the local shape of the fitness landscape. Functional knowledge about genes can be used to predict coregulation of genes. Further predictions concern the optimal feedback signals to control the genes, as well as fitness losses and differential expression after gene deletions.

5.1 Optimal regulation of stationary states

In this chapter, a quantitative model of optimal regulation is proposed: a system of regulatory variables x (for instance, gene transcript levels) affects a system of output variables y (see [15]), such as metabolic fluxes. The states of both systems are rated by a common objective function $F(x, y)$, which will be called "fitness" here. Only the "relevant" variables y which actually play a role for the fitness function will be considered. The regulators are supposed to behave optimally, that is, they always adapt their values such as to maximise the local fitness. To assume optimality is certainly an idealisation, but often used as an approximation of biological reality [24] [40] [41] [67] [105].

To illustrate the approach, let us consider how metabolic systems are controlled by differential expression of enzymes. Metabolic fluxes depend on cellular processes producing or

consuming metabolites, on environmental parameters like nutrient supply, and on parameters influencing the enzymatic activities, such as temperature. In addition, metabolism is actively controlled by regulatory processes on different time scales: while fast responses are realised by activation and inhibition of enzymes, slow adaptation is achieved by adjusting their expression. The linear influence of enzyme concentrations E_k on stationary fluxes J_i is quantified by the response coefficients matrix R^J_E. Metabolic control theory [41] [63] describes how fluxes respond to changes of enzymes, which may be caused by changes in gene expression. One may also ask the inverse question: what enzyme changes are necessary to achieve a desired metabolic behaviour, such as homoeostasis or constrained maximisation of fluxes? The answer to this question depends on (1) the control of enzyme activities on metabolism, as studied by metabolic control analysis and (2) assumptions about the objectives of the cell.

The performance of cellular subsystems can be rated by their contribution to the evolutionary fitness of the organism, that is, the expected long-term reproduction rate. In a particular environment, a few fluxes may effectively determine biomass production. For a metabolic system, we may consider a simple objective function $V(J)$ scoring only these important fluxes, and assume that there is an evolutionary tendency to maximise this function. Such an objective function was studied previously [40], notably the (mathematical) product of the two independent fluxes in a reaction system representing glucose metabolism. In the whole cell, many processes depend on common resources, so an optimal compromise must be chosen between them. The enzyme levels can adapt the metabolic system to external fluctuations, and will thereby effectively increase the fitness, but enzyme production itself consumes cellular resources, implying a negative contribution $U(E)$ to the total fitness $F(E, J) = U(E) + V(J)$. The optimal behaviour with respect to F represents a compromise between benefit and costs [95].

As a simple example (shown in Figure 5.1), let us consider a chain of two chemical reactions $S_0 \leftrightarrow S_1 \leftrightarrow S_2$ with mass-action kinetics

$$\begin{aligned} v_1 &= k_1\ E_1\ S_0 - k_{-1}\ E_1\ S_1 \\ v_2 &= k_2\ E_2\ S_1 - k_{-2}\ E_2\ S_2 \end{aligned} \tag{5.1}$$

where E_1 and E_2 denote the enzyme concentrations. At fixed concentrations S_0 and S_2, the stationary flux $J = v_1 = v_2$ reads

$$J = \frac{E_1\ E_2\ (S_0\ k_1\ k_2 - S_2\ k_{-1}\ k_{-2})}{E_1\ k_{-1} + E_2\ k_2} \tag{5.2}$$

A reasonable and frequently used ansatz for the fitness function is to use the flux itself $V(J) = J$ [40] [41] [101], while the enzyme levels are rated by a negative function

$$U(E) = -\gamma_1(E_1 + E_2) - \gamma_2(E_1 + E_2)^2 \tag{5.3}$$

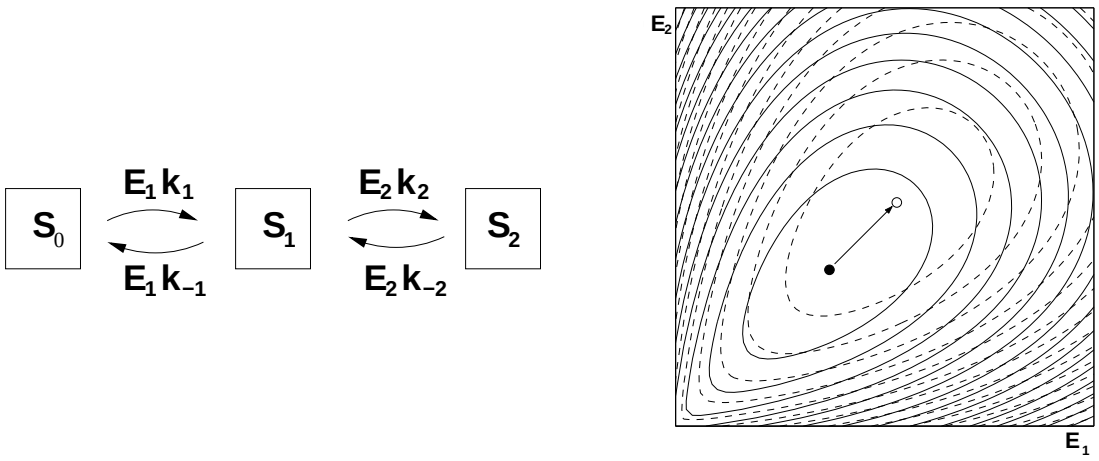


Figure 5.1: Adaptation of enzyme levels. A linear chain of two reactions (left) is controlled by the enzymes E_1 and E_2. Their performance is evaluated by a fitness function $G(E_1, E_2)$. The fitness is given by the stationary flux J minus the costs U of protein production, described by $U(E_1, E_2) = -\gamma_1(E_1+E_2)-\gamma_2(E_1+E_2)^2$. The right diagram shows the fitness landscape $G(E_1, E_2)$, for two values of the external substrate S_0 (solid and dashed contour lines, respectively). The perturbation of S_0 causes a shift of the optimum, indicated by the arrow.

The linear term describes costs per protein molecule, e.g., for the consumption of amino acids. High rates of protein synthesis require additional efforts, e.g., an increased production of ribosomes, which is punished by the quadratic term. Maximising the effective fitness

$$G(E) = F(E, J(E)) = -\gamma_1(E_1 + E_2) - \gamma_2(E_1 + E_2)^2 + J(E_1, E_2) \tag{5.4}$$

with respect to E_1 and E_2 yields unique optimal enzyme levels $(\bar{E}_1, \bar{E}_2)$. A small perturbation of the parameters, such as the concentrations of external metabolites or the rate constants, changes the fitness landscape $G(E)$ (Figure 5.1, right). The optimum is shifted, but the enzyme levels can be adapted to reach new optimal values.

This example illustrates what will now be tackled in a general way: a long-term objective is to explain correlations in genome-wide differential expression data, which would in principle require a model of the whole cell. However, local properties of the model (namely derivatives) at the initial optimal state are sufficient to predict the optimal response, provided that the perturbations are small.

5.1.1 The mathematical model

The model of optimal regulation proposed in this section describes biological regulators which control stationary states. The cell state is described by a set of output variables y that depend on regulatory variables x and on environmental parameters α. The symbols x, y, and α denote vectors. Small changes of y are expanded as

$$\Delta y(x,\alpha) \approx (R^y_x \; R^y_\alpha)\begin{pmatrix}\Delta x\\ \Delta\alpha\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\Delta x\\ \Delta\alpha\end{pmatrix}^T \begin{pmatrix}R^y_{xx} & R^y_{x\alpha}\\ R^y_{\alpha x} & R^y_{\alpha\alpha}\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta\alpha\end{pmatrix} \tag{5.5}$$

The linear influences of the regulators and the environment on y are described by the response coefficients R^y_x and R^y_α [41]. The second-order response coefficients R^y_{xx} and $R^y_{x\alpha}$ describe the quadratic effects of x and α [45]. Both x and y are rated by a fitness function $F(x,y)$, which, for simplicity's sake, is assumed to have the additive form (see also [95] [101])

$$F(x,y) = U(x) + V(y) \tag{5.6}$$

The gradient $F_y = \nabla_y F(x,y)$ will be called the marginal fitness of y, in analogy to the marginal utility defined in economics [44]. The marginal fitness $F_x = \nabla_x F(x,y)$ of x is defined accordingly. The matrices F_{xx} and F_{yy} of second-order derivatives contain the curvatures of the fitness function. If U is a sum of terms depending on the single regulators, then F_{xx} is diagonal. Sometimes an "isotropic" case will be considered, where F_{xx} is a scalar multiplied by the identity matrix I. The effective fitness $G(x,\alpha) = F(x, y(x,\alpha))$ is a function of x and α alone, with derivatives

$$\begin{aligned} G_x(x,\alpha) &= \nabla_x F(x,y(x,\alpha)) = F_x + R^{y^T}_x \; F_y \\ G_{xx}(x,\alpha) &= \nabla_x \nabla^T_x F(x,y(x,\alpha)) = F_{xx} + R^{y^T}_x \; F_{yy} \; R^y_x + T_{xx} \end{aligned} \tag{5.7}$$

as $F_{xy} = 0$ (see equation 5.6). T_{xx} represents the tensor product[1] $(T_{xx})_{ik} = (F_y)_l (R^y_{xx})^l_{ik}$. It describes an effective fitness curvature due to the cooperation of regulators, for instance gene products acting in a complex, such as in metabolic channelling [15]. Instead of assuming the cost term $U(x)$, one could describe the costly side-effects of gene expression by additional output variables y. The x-dependent fitness term F_{xx} would then reappear as a part of T_{xx}.

The optimality principle postulates that, for any given environment α, the regulators have to assume a value $\bar{x}(\alpha)$ to reach a local fitness maximum (see Figure 5.2, right). Optimality at $\bar{x}(\alpha)$ implies that G_x vanishes, so F_x and F_y are balanced according to

$$F_x = -R^{y^T}_x \; F_y \tag{5.8}$$

[1] Superscripts and subscripts of tensor symbols represent variables and derivatives, respectively. According to the sum convention, terms are summed over all indices which occur both as superscript and as subscript.

To ensure a unique local maximum, the effective fitness curvature matrix G_{xx} must have negative eigenvalues, so G_{xx} is invertible. If the number of regulators exceeds the number of output variables, then $R_x^{y^T}\ F_{yy}\ R_x^y$ in equation 5.7 has some vanishing eigenvalues, but by an appropriate choice of $U(x)$, a maximum can be ensured.

In the following, we shall study regulators in an optimal state which encounter a perturbation: two scenarios are considered, namely perturbations of y by perturbation of α, and perturbations of individual regulators x_i. In both cases, the optimal response $\mathrm{d}\bar{x}$ to maximise $\mathrm{d}F$ will be calculated in a linear approximation. Concerning the initial optimal state, some simplifying assumptions are made: locally, all values of y can be reached by an appropriate choice of x, that is, R_x^y has full row rank. This implies that the dimension of y does not exceed the dimension of x, and that $R_x^y\ F_{xx}^{-1}\ R_x^{y^T}$ is invertible. In general, the fitness function may depend on additional parameters, and the y may not be controlled independently. Formulae for these cases as well as proofs for the formulae in the following sections are given in Appendix A.

5.1.2 Adaptation to a perturbation of the output variables

Let us consider the optimal response to external perturbations of y, caused by a small change $\mathrm{d}\hat{\alpha}$. If the regulators remained constant ($\mathrm{d}x = 0$), y, F_y, and R_x^y would change by $\mathrm{d}\hat{y} = R_\alpha^y\ \mathrm{d}\hat{\alpha}$, $\mathrm{d}\hat{F}_y = F_{yy}R_\alpha^y\ \mathrm{d}\hat{\alpha}$, and $\mathrm{d}\hat{R}_x^y$, where the latter is defined by the tensor product $(\mathrm{d}\hat{R}_x^y)_i^l = (R_{x\alpha}^y)_{ik}^l\ \mathrm{d}\hat{\alpha}^k$. In this text, two sorts of differentials will be distinguished: those with a circumflex (for instance, $\mathrm{d}\hat{y}$) denote changes due to an external perturbation for fixed x, while those with a bar (e.g., $\mathrm{d}\bar{y}$) contain the additional effect of the optimal response $\mathrm{d}\bar{x}$.

To determine the response $\mathrm{d}\bar{x}$ which maximises the fitness $G(x + \mathrm{d}\bar{x}, \alpha + \mathrm{d}\hat{\alpha})$, we expand $G_x = \nabla_x G(x, \alpha)$ to first order

$$G_x(x + \mathrm{d}x, \alpha + \mathrm{d}\alpha) \approx G_x(x, \alpha) + G_{xx}(x, \alpha)\ \mathrm{d}x + G_{x\alpha}(x, \alpha)\ \mathrm{d}\alpha \tag{5.9}$$

The total differential reads

$$\mathrm{d}G_x = G_x(x + \mathrm{d}x, \alpha + \mathrm{d}\alpha) - G_x(x, \alpha) = G_{xx}\ \mathrm{d}x + G_{x\alpha}\ \mathrm{d}\alpha \tag{5.10}$$

An optimal initial state with $G_x(x, \alpha) = 0$ becomes perturbed by $\mathrm{d}\hat{\alpha}$. Without response (i.e., if $\mathrm{d}x = 0$), this would imply three changes

$$\begin{aligned} \mathrm{d}\hat{R}_x^y &= R_{x\alpha}^y \mathrm{d}\alpha \\ \mathrm{d}\hat{F}_x &= F_{xy}R_\alpha^y\ \mathrm{d}\hat{\alpha} \\ \mathrm{d}\hat{F}_y &= F_{yy}R_\alpha^y\ \mathrm{d}\hat{\alpha} \end{aligned} \tag{5.11}$$

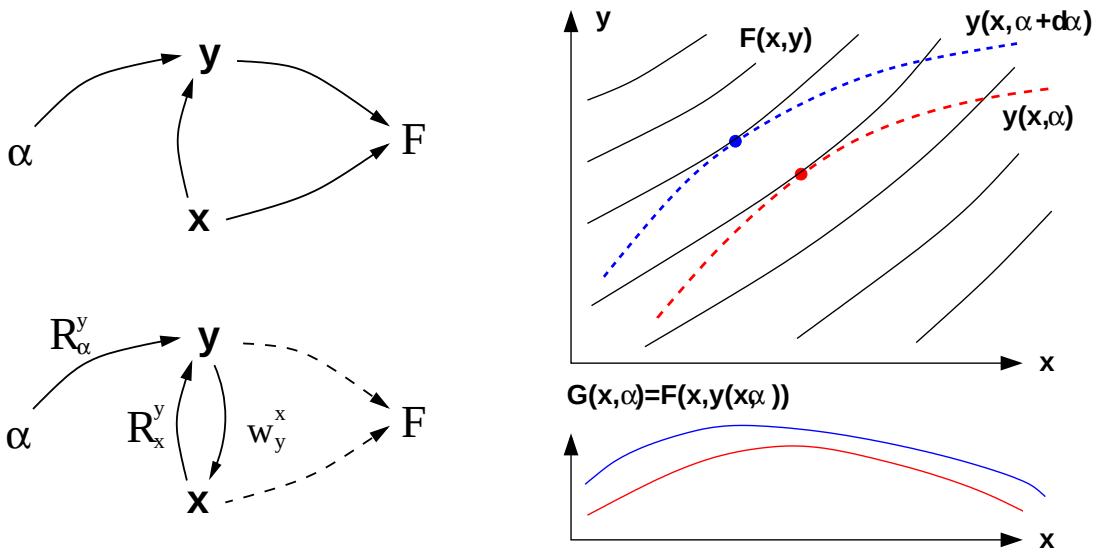


Figure 5.2: Model of optimal regulation. Top left: The output variables y is a function of the environment α and of the regulators x. The fitness function F evaluates both x and y. Bottom left: The optimal behaviour can be implemented by feedback signals between y to x (see Section 5.2.2). Optimality with respect to the fitness F (dotted arrows) is ensured by an appropriate choice of the feedback coefficients w^x_y. Right: Optimal response to a perturbation of y. A one-dimensional case is shown while in general, x, α and y are vectors. For fixed environment α, the output variable $y(x, \alpha)$ is a function of the regulator x (shown by dashed lines, for two values of α). The slope of this line is the response coefficient R^y_x. The fitness function $F(x, y)$ (shown by solid contour lines), evaluated on the curve $y(x, \alpha)$, yields the effective fitness $G(x, \alpha)$ (shown below). After a change of α, x adapts itself to reach a new optimal state (dots) maximising G.

The optimal response $\mathrm{d}\bar{x}$ must ensure that $\mathrm{d}G_x$ vanishes, so

$$\mathrm{d}\bar{x} \;=\; -G_{xx}^{-1}\,(G_{x\alpha}\,\mathrm{d}\hat{\alpha}) \tag{5.12}$$

with

$$\begin{aligned} G_{xx} &= F_{xx} + T_{xx} + R_x^{y^T} F_{yy} R_x^y + F_{xy} R_x^y + (F_{xy} R_x^y)^T \\ G_{x\alpha} &= R_x^{y^T} F_{yy} R_\alpha^y + F_{xy} R_\alpha^y + T_{x\alpha} \end{aligned} \tag{5.13}$$

The matrices $(T_{xx})_{ik} = (F_y)_l (R_{xx}^y)_{ik}^l$ and $(T_{x\alpha})_{ik} = (F_y)_l (R_{x\alpha}^y)_{ik}^l$ are calculated from the tensors R_{xx}^y and $R_{x\alpha}^y$ containing the second derivatives of $y(x,\alpha)$. Rewriting the term in brackets from equation 5.12

$$\begin{aligned} \mathrm{d}\hat{G}_x &= R_x^{y^T} (F_{yy} R_\alpha^y \mathrm{d}\hat{\alpha}) + (F_{xy} R_\alpha^y \mathrm{d}\hat{\alpha}) + (R_{x\alpha}^y \mathrm{d}\hat{\alpha})^T \; F_y \\ &= R_x^{y^T} \mathrm{d}\hat{F}_y + \mathrm{d}\hat{F}_x + \mathrm{d}\hat{R}_x^{y^T} \; F_y \end{aligned} \tag{5.14}$$

yields

$$\mathrm{d}\bar{x} \;=\; -G_{xx}^{-1} \left[R_x^{y^T} \mathrm{d}\hat{F}_y + \mathrm{d}\hat{F}_x + \mathrm{d}\hat{R}_x^{y^T} \; F_y \right] \tag{5.15}$$

Thus the regulators react to the three effects (see equation 5.11) of the perturbation. We assumed above that $F(x,y) = U(x) + V(y)$, so F_{xy} and accordingly $\mathrm{d}\hat{F}_x$ vanish. The optimal response finally reads

$$\mathrm{d}\bar{x} \;=\; -G_{xx}^{-1} \; \mathrm{d}\hat{G}_x \tag{5.16}$$

$$\text{where} \qquad G_{xx} \;=\; F_{xx} + T_{xx} + R_x^{y^T} F_{yy} R_x^y \tag{5.17}$$

$$\mathrm{d}\hat{G}_x \;=\; R_x^{y^T} \mathrm{d}\hat{F}_y + \mathrm{d}\hat{R}_x^{y^T} \; F_y \tag{5.18}$$

The symmetric matrix T_{xx} can also be incorporated into an effective fitness curvature $F_{xx}^* = F_{xx} + T_{xx}$. The terms contributing to $\mathrm{d}\hat{G}_x$ describe two effects of the perturbation, namely on the marginal fitness of y, and on the regulatory properties expressed by R_x^y. While equation 5.18 is a general result, simple consequences can be drawn if the second effect is neglected because F_y or $R_{x\alpha}^y$ is sufficiently small. With this simplification, the remaining optimal response reads

$$\mathrm{d}\bar{x} = -(F_{xx} + T_{xx} + R_x^{y^T} F_{yy} R_x^y)^{-1} \; R_x^{y^T} \; F_{yy} \; R_\alpha^y \; \mathrm{d}\hat{\alpha} \tag{5.19}$$

Note that only the second derivatives of the fitness appear in the formula, because the first derivatives are initially balanced (see equation 5.8).

Instead of being neglected, the second term in equation 5.18 can also be incorporated into the first one if the perturbation leaves the normalised response coefficients $x_k/y_i (R_x^y)_{ik}$

constant. Then we get

$$\begin{aligned} \frac{(\mathrm{d}R_x^y)_{ik}}{(R_x^y)_{ik}} &= \frac{\mathrm{d}y_i}{y_i} \\ \Rightarrow \quad \mathrm{d}R_x^y &= \mathrm{dg}(\mathrm{d}y)\ \mathrm{dg}(y)^{-1}\ R_x^y \end{aligned} \tag{5.20}$$

The last term of equation 5.18 can therefore be rewritten as

$$R_x^{y^T}\ \mathrm{dg}(F_y)\ \mathrm{dg}(y)^{-1}\mathrm{d}\hat{y} \tag{5.21}$$

and be incorporated into the first term: bearing in mind that $\hat{F}_y = F_{yy}\mathrm{d}\hat{y}$, we obtain

$$\mathrm{d}\hat{G}_x = R_x^{y^T}\ (F_{yy} + \mathrm{dg}(F_y)\ \mathrm{dg}(y)^{-1})\ \mathrm{d}\hat{y} = R_x^{y^T}\ F_{yy}^*\mathrm{d}\hat{y} \tag{5.22}$$

Effectively, the second term has disappeared, while F_{yy}^* contains an additional contribution $\mathrm{dg}(F_y)\ \mathrm{dg}(y)^{-1}$. Is it a reasonable assumption that the normalised response coefficients are not affected by perturbations? For a linear reaction chain with linear kinetics (see [41]), the normalised response coefficients do not depend on the substrate concentration, while they do depend on the enzyme parameters. Thus the assumption holds for a perturbation of the substrate, but not for a perturbation of the enzyme parameters.

5.1.3 Obtaining a change of the output variables

Let us now consider a different setting where the regulators must achieve a fixed change $\mathrm{d}y$. Under the constraint that $\mathrm{d}y = R_x^y\ \mathrm{d}\bar{x}$, the fitness is maximised by (proof: Appendix A.2)

$$\mathrm{d}\bar{x} = F_{xx}^{-1}R_x^{y^T}(R_x^y F_{xx}^{-1}R_x^{y^T})^{-1}\ \mathrm{d}y \tag{5.23}$$

In the isotropic case, this reduces to $\mathrm{d}\bar{x} = R_x^{y^+}\ \mathrm{d}y$, with the pseudoinverse[2] of R_x^y. If the fitness term U rates the regulators separately, then F_{xx} is diagonal, and the diagonal elements of the first term F_{xx}^{-1} appear as weights in formula 5.23: a large negative curvature leads to a weak response of the respective regulator $\mathrm{d}\bar{x}_i$. For reasons of consistence, equation 5.23 must also hold for any optimal contribution $\mathrm{d}y = \mathrm{d}\bar{y} - \mathrm{d}\hat{y}$ after a perturbations $\mathrm{d}\hat{\alpha}$. Thus we obtain the central result that, for isotropic F_{xx}, any optimal expression profile is a linear combination of regulatory profiles, i.e., the rows of R_x^y. On the other hand, if y must keep its original value despite a perturbation $\mathrm{d}\hat{\alpha}$, then $\mathrm{d}\bar{y} = R_x^y\ \mathrm{d}\bar{x} + R_\alpha^y\ \mathrm{d}\hat{\alpha}$ has to vanish, so we set $\mathrm{d}y = -R_\alpha^y\ \mathrm{d}\hat{\alpha}$.

[2]The pseudoinverse of a matrix A is defined as $A^+ \equiv (A^T\ A)^{-1}A^T$.

5.1.4 Adaptation to a perturbation of individual regulators

Besides perturbations of the output variables y, we can study perturbations of individual regulators x. In the case of gene expression, such perturbations may be realised by gene deletions [51] or RNA interference [29] or may result from hereditary enzyme deficiencies. In the model, one regulator is driven away from the local optimum of the fitness landscape $G(x, \alpha)$, but the others can compensate for the loss. Let us assume that regulator x_i is changed[3] by a fixed value $\mathrm{d}\hat{x}_i$, that is, $\mathrm{d}\hat{x} = (0 \,..\, 0 \; \mathrm{d}\hat{x}_i \; 0 \,..\, 0)^T$. The optimal response of the other regulators reads (proof: see Appendix A.3)

$$\mathrm{d}\bar{x} = G_{xx}^{-1} \frac{1}{(G_{xx}^{-1})_{ii}} \; \mathrm{d}\hat{x} \tag{5.24}$$

The small perturbation of a single gene leads to a fitness loss

$$\mathrm{d}^2 G = \frac{1}{2} \; \mathrm{d}\bar{x}^T \; G_{xx} \; \mathrm{d}\bar{x} = \frac{1}{2} \; \frac{(\mathrm{d}\hat{x}_i)^2}{(G_{xx}^{-1})_{ii}} \tag{5.25}$$

Small diagonal elements of G_{xx}^{-1} imply large fitness losses and may indicate essential genes.

Depending on the curvatures of the effective fitness landscape, gene pairs will either show coregulation or anti-coregulation as one of the genes is deleted (see Figure 5.3). Both kinds of behaviour are even possible for genes exerting the same first-order control, described by R_x^y. Cooperating genes may also be coregulated on an evolutionary time-scale, by mutations: if one gene is deleted, a deletion of the second one should become an advantage. Thus pairs of cooperating genes may appear in phylogenetic profiles [91], while pairs of genes compensating for each other should show phylogenetic anticorrelation [86].

5.2 Feedback signals and the value of regulators

5.2.1 A cascade of responses

Near a fitness maximum, a regulatory system x buffers fitness fluctuations, in analogy to le Châtelier's principle. This buffering can be described by a cascade of responses. Let us recall equation 5.16: if the marginal effective fitness of the regulators is perturbed by an amount $\mathrm{d}\hat{G}_x$, the matrix G_{xx}^{-1} describes how this perturbation becomes distributed over the whole system. If the fitness curvature with respect to x is high, that is, if $F_{xx}^{-1} R_x^{y^T} F_{yy} R_x^y$

[3] Alternatively, the perturbation can be modelled as a marginal fitness change $\mathrm{d}\hat{G}_x = F_{x\beta} \; \mathrm{d}\hat{\beta}$ due to an additional parameter β in the fitness $G(x, \alpha, \beta)$. The optimal response then reads $\mathrm{d}\bar{x} = -G_{xx}^{-1} \; \mathrm{d}\hat{G}_x$

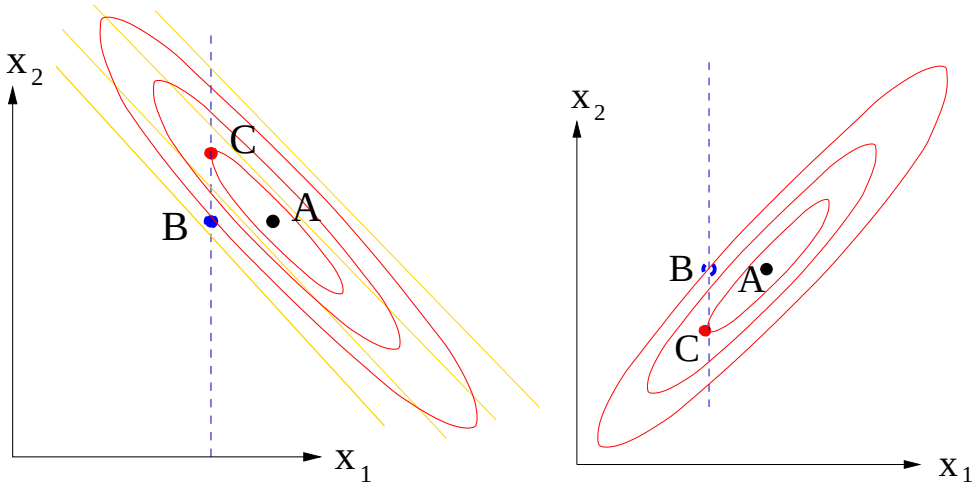


Figure 5.3: Interplay between regulators in the fitness landscape $G(x, \alpha)$ for fixed environment α. Left: The fitness G with respect to two genes x_1 and x_2 has elliptic contour lines, with the optimum in the centre (A). Constraining x_1 to a smaller value (dashed line) would decrease the fitness (B). An activation of x_2 damps the fitness loss (C). The fitness landscape shown may result from a gene duplication: if both genes exert the same influence on y and if $F_{xx} = 0$, then the maximum of the fitness (diagonal contour lines) is non-unique. A finite term F_{xx} regularises the effective fitness function G (i.e. it causes all eigenvalues of G_{xx} to be nonzero), which leads to the elliptic contour lines. The term F_{xx} can be caused by a nonlinear effect in the cost term $U(x)$: for example, it might be more costly to synthesise two isoenzymes than to synthesise one enzyme with a given function. Right: Cooperation can be induced by second-order response terms R^y_{xx} contributing to F^*_{xx}. If genes x_1 and x_2 are both necessary for the same process, they tend to be coregulated.

has small absolute eigenvalues, then G_{xx}^{-1} can be expanded into a power series (proof: see Appendix A.4, also compare [39])

$$G_{xx}^{-1} = (1 + F_{xx}^{-1} R_x^{y^T} F_{yy} R_x^y)^{-1} F_{xx}^{-1} = \sum_{n=0}^{\infty} (-F_{xx}^{-1} R_x^{y^T} F_{yy} R_x^y)^n \; F_{xx}^{-1} \tag{5.26}$$

The series describes superposed responses of different order: an immediate response to the perturbation, which may have unfavourable side-effects, a response to these effects, and so on. The complete response represents a systemic compromise between all effects of the regulators. It has to be stressed that the cascade does not describe time-dependent behaviour.

On the other hand, if $F_{xx} + T_{xx}$ is small ($F_{xx} + T_{xx} \to 0$), equation 5.19 yields the optimal response (proof: see Appendix A.5)

$$\mathrm{d}\bar{x} \;\approx\; -R_x^{y^+} F_{yy}^{-1} \; \mathrm{d}\hat{F}_y \tag{5.27}$$

as it would result from optimising first $\mathrm{d}y$, and then $\mathrm{d}x$.

5.2.2 Optimal control realised by feedback

Biological regulators often receive signals from the processes to be regulated: this phenomenon is known as feedback. Gene expression, for instance, is controlled by transcription factors that provide information about the cell status. It is a basic assumption of the present analysis that during evolution, adaptation mechanisms for coping with variable environmental conditions have developed and can be described by optimality principles. This assumption is now used for describing feedback systems: the objective is to derive a feedback system that realises the optimal behaviour of regulators defined above.

Let us consider a system of interacting regulators x and cell variables y in a stationary environment α: if α is replaced by $\alpha + \mathrm{d}\alpha$, then the stationary state values of x and y respond by changes (see Figure 5.2, left bottom) fulfilling

$$\begin{aligned} \mathrm{d}x &= w_y^x \; \mathrm{d}y \\ \mathrm{d}y &= R_x^y \; \mathrm{d}x + R_\alpha^y \; \mathrm{d}\alpha \end{aligned} \tag{5.28}$$

The linear coefficients w_y^x represent the partial derivatives of a (possibly nonlinear) feedback function. For example, the activity of an enzyme can be affected by a metabolite concentration via allosteric regulation (described by w_y^x). At steady state, this concentration, in turn, is a function of all enzyme concentrations in the reaction network (described by R_x^y). A small perturbation $\mathrm{d}\alpha$ provokes a response

$$\mathrm{d}x \;=\; (1 - w_y^x R_x^y)^{-1} (w_y^x \; R_\alpha^y) \; \mathrm{d}\alpha \tag{5.29}$$

How can we ensure that this response maximises a given fitness function? If the second term in equation 5.18, describing a perturbation of the response coefficients, is neglected, the following relation holds between the optimal response $\mathrm{d}\bar{x}$ and the resulting change $\mathrm{d}\bar{y}$ (proof: Appendix A.7)

$$\begin{aligned} \mathrm{d}\bar{x} &= -F_{xx}^{-1} R_x^{y^T}\ \mathrm{d}\bar{F}_y & (5.30) \\ &= -F_{xx}^{-1} R_x^{y^T} F_{yy} \mathrm{d}\bar{y} & (5.31) \end{aligned}$$

Comparing equations 5.28 and 5.31 yields the optimal the feedback coefficients

$$w_y^x = -F_{xx}^{-1} R_x^{y^T} F_{yy} \tag{5.32}$$

The feedback to a regulator depends on the regulator's influences w_y^x, weighted by the fitness curvatures. An output variable with large negative fitness curvature will send strong feedback signals, a regulator with large negative fitness curvature will receive weak signals. So, feedback signals represent the most important variables and affects the most responsible regulators. Let us consider again allosteric control in metabolism: if homoeostasis in metabolism is to be ensured, equation 5.32 predicts feedback from metabolites to those reactions exerting a considerable control on the metabolite. If the curvatures F_{xx} and F_{yy} are negative and the reaction exerts a positive control on the metabolite, a negative feedback is predicted.

5.2.3 The value of regulators

What quantitative advantage does a regulatory system provide to the organism? To answer this question, we have to refer to an specific ensemble of external conditions: if the perturbations $\mathrm{d}\alpha$ are small and normally distributed with mean $\langle \mathrm{d}\alpha \rangle = 0$ and covariance matrix $\mathrm{cov}(\mathrm{d}\alpha) = \langle \mathrm{d}\alpha\ \mathrm{d}\alpha^T \rangle$, the presence of the regulating system raises the fitness, on average, by (proof: see Appendix A.6)

$$\langle \bar{G} - \hat{G} \rangle = -\frac{1}{2}\ Tr(G_{\alpha x}\ G_{xx}^{-1}\ G_{x\alpha}\ \mathrm{cov}(\mathrm{d}\alpha)) \tag{5.33}$$

As G_{xx} has no positive eigenvalues, the value $\langle \bar{G} - \hat{G} \rangle$ of the regulatory system is nonnegative. The name "value" has been chosen in analogy to the value of information. Evolution is likely to develop regulators of high value: if the very presence of a regulatory system involves additional costs, it should only be maintained if its value exceeds the costs. Like the information value, which depends on the presence of other information sources, the value of regulators may be influenced by the presence of other regulators. For instance, adding copies of existing regulators to the system will not yield much additional fitness.

We may also consider partially informed regulators which respond to signals $\sigma(\alpha)$ providing noisy information about the perturbation via a conditional probability $p(\alpha|\sigma)$. If the regulators can only sense σ, they must have to optimise the expected fitness given the signal σ, $G^* = \int G(x,\alpha)p(\alpha|\sigma)d\alpha$, which can be regarded as a new, effective fitness function. For this effective fitness, the value of regulation equals the value of information.

5.3 Predictions for gene expression patterns

5.3.1 Correlation between functionally related genes

Coregulation of genes is often quantified by the linear correlation that is, by the covariance between gene profiles, normalised by the square root of their variances. Given the covariance $\mathrm{cov}(\mathrm{d}\hat{F}_y)$ between the marginal fitness perturbations of y, the covariances between the responses $\mathrm{d}\bar{x}_i$ read (see equation 5.30)

$$\begin{aligned} \mathrm{cov}(\mathrm{d}\bar{x}) &= G_{xx}^{-1}\ R_x^{y^T}\mathrm{cov}(\mathrm{d}\hat{F}_y)\ R_x^y\ G_{xx}^{-1} \\ &= F_{xx}^{-1}\ R_x^{y^T}\mathrm{cov}(\mathrm{d}\bar{F}_y)\ R_x^y\ F_{xx}^{-1} \end{aligned} \tag{5.34}$$

For a strong isotropic fitness curvature ($F_{xx} \to -\infty$), this yields, in first order

$$\mathrm{cov}(\mathrm{d}\bar{x}) \propto R_x^{y^T}\mathrm{cov}(\mathrm{d}\hat{F}_y)R_x^y \tag{5.35}$$

In this approximation, two genes are correlated if they have strong effects on the same variables, or on variables with large common fluctuations of the marginal fitness. Accordingly, cooperating proteins are likely to show correlated expression, as it was empirically found for interacting proteins [36], permanent protein complexes [62], and subsets of cooperating enzymes [104].

5.3.2 Correlated expression of interacting proteins

Pairs of interacting proteins tend to show correlated expression, and an increased correlation is also observed for pairs of essential proteins. Figure 5.4 shows correlations of expression profiles from the cell stress data Gasch et al. [32]. The diagrams on the left show correlation histograms for pairs of interacting proteins, determined from a yeast two-hybrid assay [112], for pairs of essential genes according to MIPS [82] and the experiment by Giaever et al. [35], and for pairs of interacting essential proteins. Compared to randomly chosen gene pairs (dashed lines), correlations of the selected pairs are shifted towards positive values.

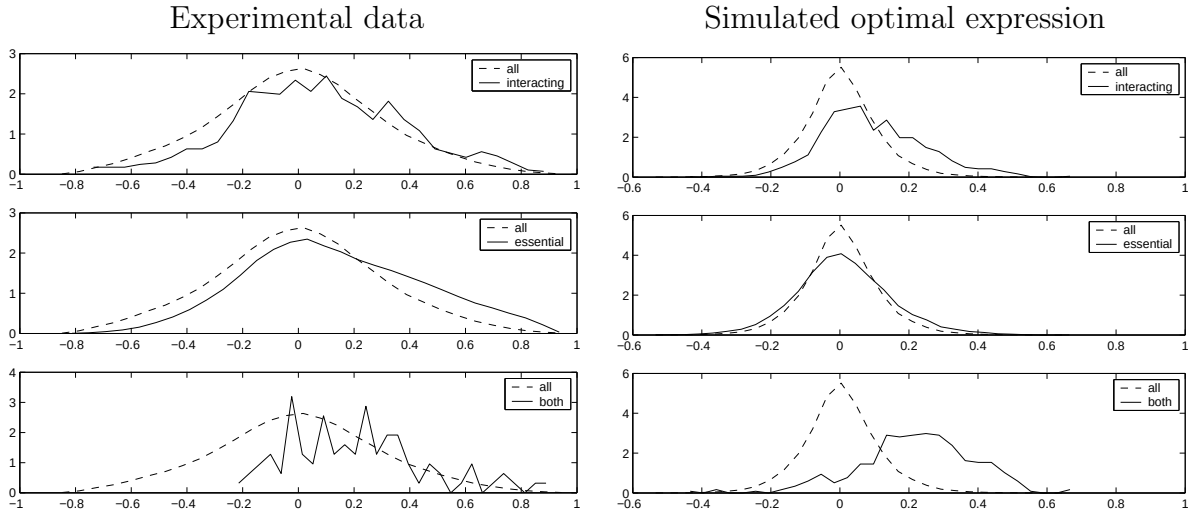


Figure 5.4: Correlated expression of interacting proteins and pairs of essential genes. Left: Histogram of correlations, calculated from the stress response data [32]. The diagrams show correlations for interaction pairs from [112] (solid line, top), essential genes according to the MIPS database ([82]) and [35] (solid line, centre), and pairs of interacting essential proteins (solid line, bottom). Compared to randomly chosen gene pairs (dashed line), the (normalised) histograms are shifted towards positive values. Right: Correlations between optimal regulation patterns show a similar qualitative behaviour. The expression data were simulated for a system with randomly chosen response coefficients, but accounting for protein complexes (see text). For interacting proteins, the histogram is shifted, while for essential genes, it becomes broader.

A similar qualitative behaviour occurs in simulations of optimal expression patterns, based on a model with randomly chosen response coefficients: the response coefficients matrix consists of two parts $R^y_x = R^{y*}_x + R^y_c R^c_x$ where R^{y*}_x describes the direct influence of proteins, while $R^y_c R^c_x$ describes the influence of complexes built from the proteins[4]. Covariances between responses $\mathrm{d}\bar{x}$ were calculated[5] according to equation 5.34. The correlations were calculated from the covariance matrix according to

$$\mathrm{cor}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{(\mathrm{cov}(x, x) + \rho)(\mathrm{cov}(y, y) + \rho)}} \tag{5.36}$$

to account for small additive noise with variance ρ. Genes from the same complex were assumed to represent interaction pairs, while genes with low diagonal elements $(G^{-1}_{xx})_{ii}$ (compare equation 5.25) were considered essential. Histograms of the simulated correlations are shown on the right side of Figure 5.4: the correlations for interacting gene pairs are shifted towards positive values, while for essential genes, the histogram becomes broadened.

5.3.3 Symmetric compensation for deletions

Let us consider a deletion experiment in which, in the i^{th} sample, gene x_i (logarithmic expression value) is downregulated by $\mathrm{d}\hat{x}_i$. According to the equation 5.24, the expression matrix X with the experiments in the rows should be decomposable into

$$X = G^{-1}_{xx}\ D \tag{5.37}$$

where D is diagonal. The symmetry of G^{-1}_{xx} implies a symmetric relation between the genes: if the loss of gene A leads to an activation of gene B, gene A should also be activated after the loss of gene B. Matrices derived from experimental data according to equation 5.37 were tested for their symmetry (see Figure 5.5). Ideker et al. [61] studied deletions of enzymes in the galactose pathway: the estimate[6] of G^{-1}_{xx} according to equation 5.37

[4] R^{y*}_x and R^y_c are sparse, while R^c_x is block-diagonal, relating the genes to complexes of sizes between 1 and 15. Each gene participates exactly in one complex. For all these matrices, nonzero elements were drawn from a normal distributions with standard deviation 1 and positive mean. Both $R^y_c R^c_x$ and R^{y*}_x were normalised by the root mean square of their elements, to give them equal weight in R^y_x.

[5] For $-F_{yy}$ and C, symmetric matrices with log-normal eigenvalues were used, and an isotropic F_{xx} was chosen.

[6] The column and row means of the whole data set (log10 expression ratios) were adjusted to zero, and a submatrix related to the genes GAL1, GAL2, GAL3, GAL4, GAL7, and GAL10 and the respective knock-out mutants was chosen. I calculated the difference matrix X between the respective "+gal" and "-gal" samples and determined a diagonal matrix D such that the mean squares for the rows of XD^{-1} were similar to those of the columns. To do so, the matrix rows were iteratively scaled by the ratio between the sum of squares within columns and within rows.

shows a strong symmetric part. Hughes et al. [51] deleted 248 genes[7] of various functions: here G_{xx}^{-1} shows only weak symmetry. The reason may be that many genes knocked out were transcription factors of various functions, so we expect weak off-diagonal elements in G_{xx}. However, for metabolic genes, the matrix still contains a significant symmetric part[8]. Thus reciprocal compensation is found within the galactose pathway, but much less between different functional subsystems of the cell. It is questionable whether a gene deletion can be treated as a small perturbation. In some cases, this may indeed be the case, notably if the effects of the deletion are sufficiently buffered by the adaptation of other genes.

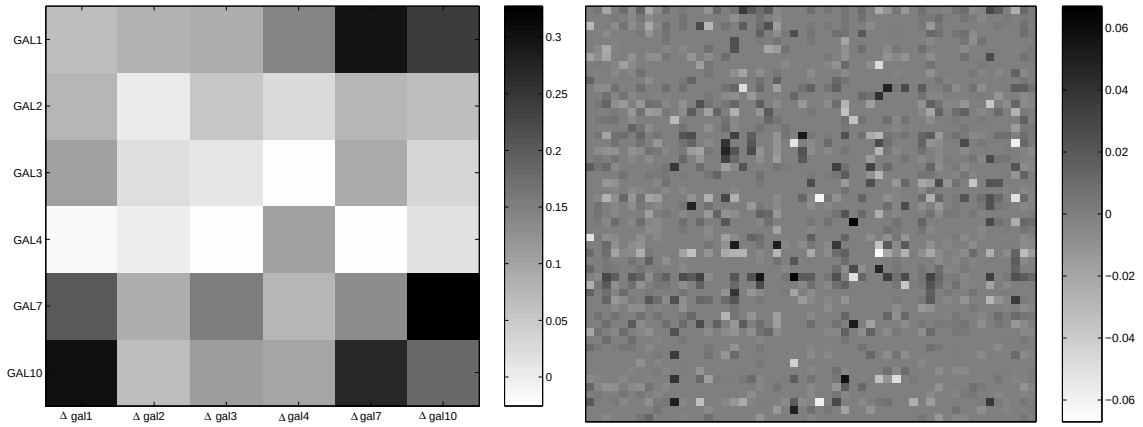


Figure 5.5: Symmetric response to deletions. Expression matrices from deletion experiments were studied: the columns correspond to deleted genes, while the rows correspond to the measurement of the same genes. According to equation 5.37, the expression matrices were decomposed into a diagonal matrix and an estimate of the inverse fitness curvature matrix G_{xx}^{-1}. Symmetry of the reconstructed G_{xx}^{-1} was tested for two data sets. Left: The matrix G_{xx}^{-1} extracted from Ideker et al. [61] shows a strong symmetric part. Right: Matrix extracted from Hughes et al. [51]. The symmetric part is weak but significant (see text).

[7] Some genes were represented by more than one ORF

[8] Only the 53 genes annotated with an EC number, according to KEGG [64], were chosen. Values where the estimated error of log ratios exceeded 2 or two times the absolute value were neglected, and variance stabilisation [50] was applied to the remaining values. For determining D, the neglected values were formally set to 0. The symmetry of the resulting matrix is weak. To decide whether the symmetric part was still significant, I calculated the standard deviations of the symmetric and antisymmetric parts (for the "good" off-diagonal entries). The ratio of about 1.7 has a p-value of about 0.01 as calculated by a permutation test where the order of the matrix rows was randomised 500 times.

5.3.4 Growth of deletion mutants

After a change of the environmental conditions, some gene products may become especially important for surviving. They should be activated, and their loss by a deletion should have a strong impact on the growth rate, while the loss of a dispensable gene should play a minor role. Thus a relation between expression data and the growth rates in deletion experiments may be hypothesised. Giaever et al. [35] studied the growth rate of yeast deletion mutants under different experimental conditions and compared the results to expression data for the same conditions: except for the growth on galactose, their experiments gave almost almost evidence for such a relation, but this was seen as a surprise. The model of optimal regulation, though, supports the initial hypothesis, predicting a quantitative relation between the data from expression and deletion experiments.

How does a deletion influence the growth rate under different conditions? A small environmental perturbation $\Delta\alpha$ and a small regulatory change Δx lead to a fitness change

$$\begin{aligned} \Delta G &\approx G_x^T \Delta x + G_\alpha^T \Delta\alpha + \frac{1}{2}(\Delta x^T G_{xx} \Delta x + \Delta\alpha^T G_{\alpha\alpha} \Delta\alpha + 2\Delta x^T G_{x\alpha} \Delta\alpha) \\ &= \left[G_x^T \Delta x + \frac{1}{2}\Delta x^T G_{xx} \Delta x\right] + \left[G_\alpha^T \Delta\alpha + \frac{1}{2}\Delta\alpha^T G_{\alpha\alpha} \Delta\alpha\right] + \Delta x^T G_{x\alpha} \Delta\alpha \quad (5.38) \end{aligned}$$

The fitness change consists of three terms, one caused by the deletion, one due to the changed conditions, and one representing the interaction between both effects, which should manifest themselves in the data matrix. If the rows and columns of the data matrix are centred, the matrix will basically represent the interaction term. According to equation 5.24, the optimal response to a deletion $\Delta\hat{x}_i$ is $\Delta\bar{x} = G_{xx}^{-1} \frac{1}{(G_{xx}^{-1})_{ii}} \Delta\hat{x}$. Inserting this into the interaction term from 5.38 yields the fitness loss

$$\frac{1}{(G_{xx}^{-1})_{ii}} \Delta\hat{x}_i^T G_{xx}^{-1} G_{x\alpha} \Delta\alpha \quad (5.39)$$

For each gene i, this term is proportional to the differential expression under the different conditions described by $\Delta\alpha$ (see equation 5.12).

5.4 Examples

Simple metabolic network. Optimal regulation of metabolic fluxes is illustrated in Figure 5.6 for a simple network of irreversible reactions, containing 8 metabolites, four of which are external. Each reaction J_i is catalyzed by an enzyme E_i. A value of 1 was chosen for the elasticity between a reaction and its substrate, while all other elasticities vanish. The fitness function depends on the fluxes J_1, J_2, and J_6, and on the enzyme

concentrations. The fluxes J_1, J_2, and J_6 are evaluated by a fitness function with the local curvature matrix $V_{JJ} = -I$. A function U with equal curvatures $U_{EE} = -I$ describes the fitness contribution of the enzyme levels E_i. The slopes of the fitness do not appear in the formulae and thus need not be specified. For illustration, specific external perturbations decrease one of the fluxes J_1, J_2, and J_6, while leaving the others unchanged. The two scenarios from sections 5.1.2 and 5.1.4 are considered: the diagrams in the upper box show the optimal response (according to equation 5.19) to a specific perturbation of J_1, J_2, or J_6, respectively. In each diagram in the lower box, one of the enzymes E_1, E_3, E_6, E_9 is inhibited, that is, constrained to a lower value. The remaining enzymes adapt themselves optimally, according to equation 5.24. For both scenarios, all enzymes respond in a coordinated way: fluxes in the whole system are redirected to increase the perturbed flux, and thus to damp the perturbation.

Glycolysis in yeast. For a second example, empirical control coefficients were taken from the glycolysis model of Hynne et al. [52]. The model describes chemical reactions related to glycolysis with kinetic parameters estimated for the onset of glycolytic oscillations. The fitness function was chosen to evaluate the influx of Glucose and the concentrations of ATP and NAD^+. Like in the previous example, the control coefficients were calculated from the stoichiometric matrix and the elasticities, and a simple fitness function ($F_{YY} = -I$, $F_{EE} = -I$) was assumed. Figure 5.7 shows optimal expression patterns after perturbations of Glucose influx, ATP and NAD^+, calculated according to equation 5.19. The responses damp the effects of the perturbation and thus contribute to homoeostasis: to increase the flux of Glucose, hexokinase is upregulated. To increase the ATP production, phosphofructokinase, is activated while other ATP-consuming reactions and the storage of ATP are inhibited.

Growth, damage repair, and energy. The optimal feedback (see section 5.2.2) is illustrated in Figure 5.8 by a schematic model describing the balance between cell growth, damage repair, and energy production. The effect of heat stress on growth and energy production has been studied experimentally in [81]. The three cell variables represent growth rate, cell damage, and energy status, and each of them is controlled by a regulator. Cell damage and energy status are influenced by external conditions, namely temperature and food supply. As above, the fitness curvatures with respect to the regulators read $F_{xx} = -I$. F_{yy} contains an additional off-diagonal element to punish growth in the presence of damage:

$$F_{yy} = -\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

The optimal feedback signals, determined by equation 5.32, are shown as a network on the right hand side of Figure 5.8.

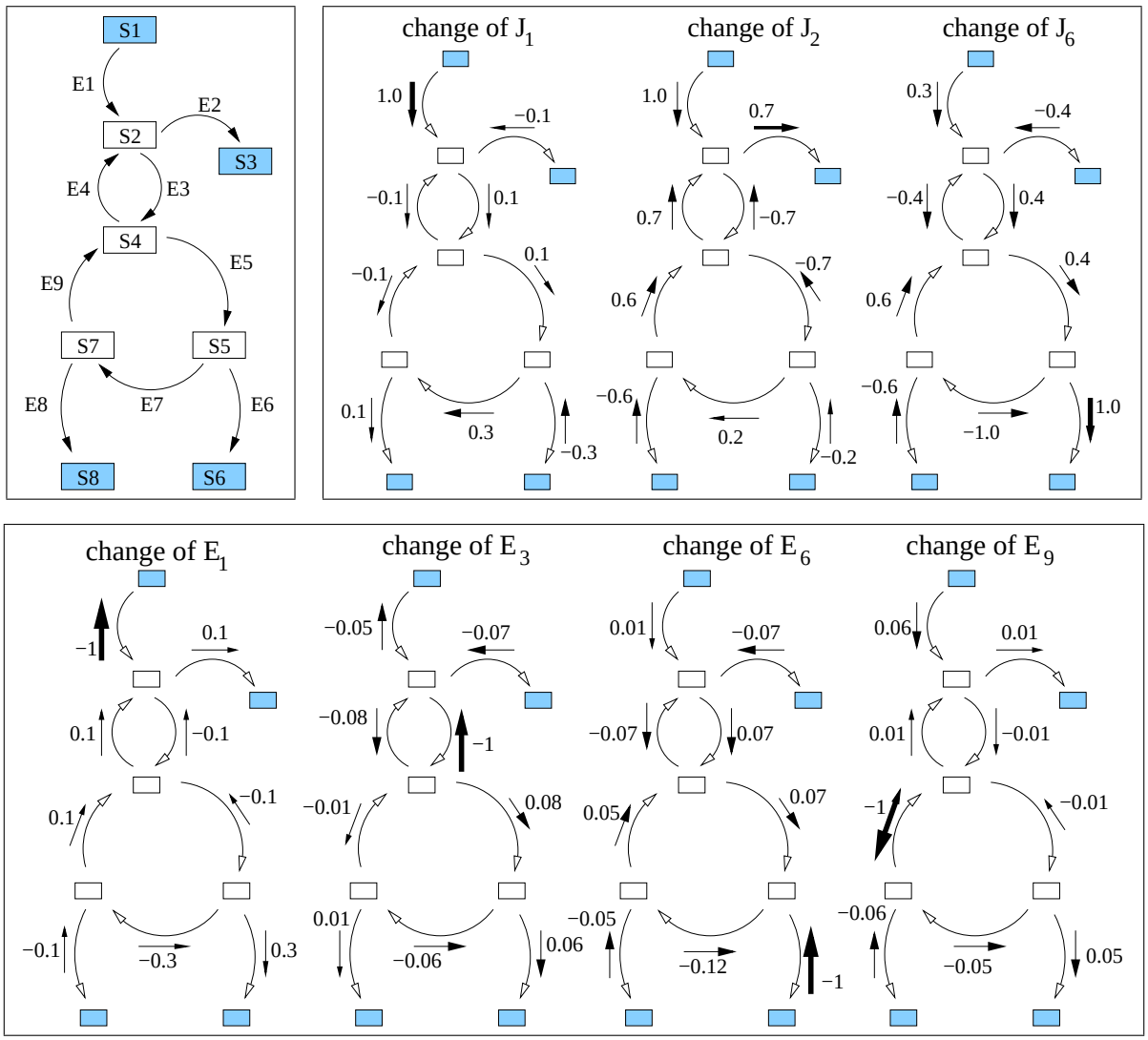


Figure 5.6: Optimal regulation of a simple network of irreversible reactions (top left box) containing 8 metabolites (shown as rectangles). The metabolites S_1, S_3, S_6, and S_8 (shaded) are considered external. Each reaction J_i is catalyzed by an enzyme (regulator) E_i. The fitness function depends on the fluxes J_1, J_2, and J_6, and on the enzyme concentrations. Top right: Each diagram shows the optimal response to a specific perturbation of J_1, J_2, or J_6, respectively. The effect of the adaptation is shown by the arrows: arrowheads indicate the direction of the immediate flux change $R_E^J \, \mathrm{d}\bar{E}$ caused by regulation. The numbers denote the adaptations $\mathrm{d}\bar{E}_i$, normalised to $\max(|\mathrm{d}\bar{E}_i|) = 1$ for each diagram. Bottom: In each diagram, one of the enzymes E_1, E_3, E_6, E_9 (indicated by a thick arrow) is inhibited, i.e., constrained to a lower value. The remaining enzymes adapt themselves and damp the perturbation.

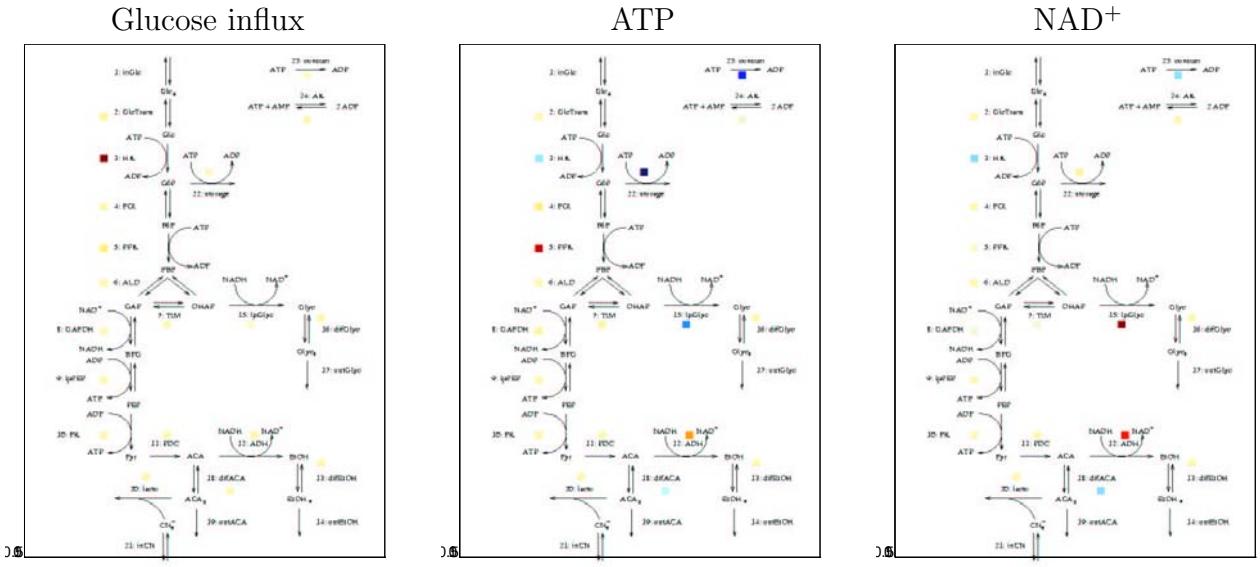


Figure 5.7: Optimal control of Glucose, ATP and NAD^+ based on the yeast glycolysis model of Hynne et al. [52]. The model scheme was taken from the original publication. The fitness function evaluates three cell variables, namely the influx of Glucose and the concentrations of ATP and NAD^+. Each diagram shows the optimal response to a forced decrease of one of these variables. The differential expression values are shown by colours: blue and red represent negative and positive values, respectively.

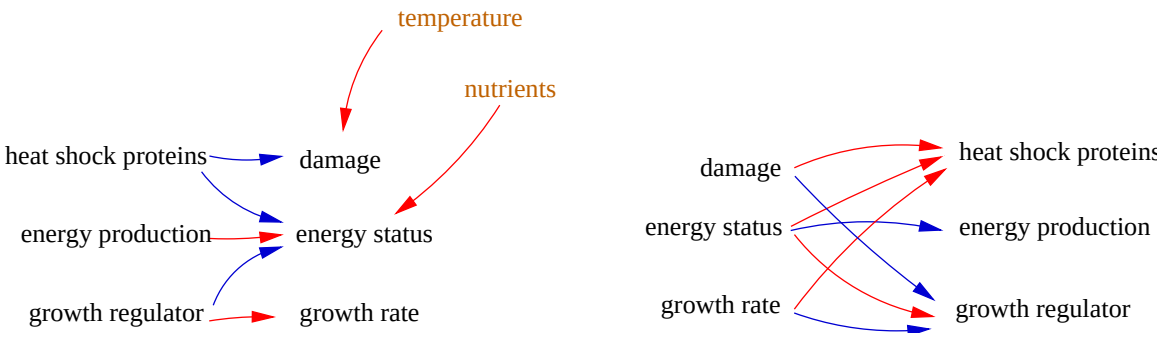


Figure 5.8: Regulation of growth, damage repair, and energy production. Left: The response coefficients (elements of R_x^y and R_α^y) describe how the regulators (heat shock proteins, energy production, and growth regulation) and the environmental parameters (temperature and nutrients) influence the output variables. They are depicted by arrows: red and blue arrows indicate the values +1 and -1, respectively. Heat shock proteins reduce cell damage, growth regulators increase cell growth, and both actions require energy and thus lower the energy status of the cell. Right: Optimal response to perturbations can be ensured by a feedback network with coefficients w_y^x, as determined by equation 5.32. Here red and blue arrows denote positive and negative values, respectively: cell damage induces the repair mechanism and downregulates growth. A forced increase of the growth rate has the same effects, while a high energy status induces both repair and growth and decreases energy production.

5.5 Discussion

The present analysis of optimal differential expression is based on the assumption that living organisms do not only have certain optimality properties in their basic (healthy) state but also respond to perturbations in an optimal way. They are assumed to reach a state that is optimal under the new conditions, thus partly compensating for the impairment due to the perturbation. A distinction has been made between regulatory variables and output variables. Specific examples could be the concentrations of gene products (e.g. enzymes) and metabolic fluxes. Accordingly, gene expression was used as a running example. However, the proposed model is not limited to gene expression: it may be applied to the design of various regulatory systems on different time scales, such as enzyme kinetics, allosteric control, adaptation of receptors, and even evolution of enzyme properties. Moreover, it is applicable to systems of any size. A promising application of the method is the analysis of DNA microarray experiments where healthy states are compared with perturbed (e.g. diseased) states. As an outcome of evolutionary optimisation, a gene's optimal expression profile depends on the gene's capacity to influence the cell state, and as a consequence, differential expression patterns tend to portray functional structures of the cell. This correspondence between expression and function will be further studied in chapter 9.

Fitness function. The success of the method largely depends on the choice of the objective function. This is a general problem in the modelling of optimal properties of living organisms [39] [40]. In any case, the biological costs for the regulatory variables should be taken into account. This can be done (and has been done here) by including a negative cost term into the fitness function. In unbranched enzymatic chains, equating the fitness function with the metabolic flux minus a linear combination of enzyme concentrations is a reasonable choice [95]. Another possibility is to use a side constraint related to the costs [39]. In either case, biological behaviour is regarded as the solution to an economical problem [95], namely to choose optimal compromises between possible actions which maximise a utility function [44]. A related optimisation problem also appears in biotechnology, namely to increase the yield of a metabolite by the modification of single genes: the costs depend on the number of genes to be engineered, so only the genes which exert the highest control on the respective metabolite will be modified. On the contrary, the present model, in which the number of responding genes does not play a role, claims that all genes should be adapted, but those with the highest control should be adapted most strongly.

Validity of the optimality assumption. The present model of regulation relies entirely on an optimality principle. However, it is not clear to which extent biological regulators realise an optimal behaviour. Segrè et al. [105] found evidence for non-optimal adaptation of metabolic fluxes after gene deletions in E. coli, but their ansatz for the

fitness function does not account for costs of the expression machinery, so it cannot be compared directly to the approach of this work. On the other hand, experiments [51] [35] have shown that gene deletions can increase the growth rate of yeast. Acting optimally to maximise its growth rate, the cell would anticipate any possible advantageous deletion by downregulating the respective genes, so no further increase would be possible. The present theory may fail here for several reasons: either no steady-state function is optimised at all by the cells, or the growth rate is not the (only) target for optimisation. Moreover, the experimental conditions possibly did not reflect the typical environment during evolution, or the deletions had side effects that could not be achieved by a change in expression alone. After all, one cannot hope to deduce all biological behaviour from an optimality principle, and it is an open question in which cases optimality assumptions are valid. At least, two conditions should be met: the experiment must probe the cell with physiological conditions to which the system has become accustomed during evolution, and for the present analysis, the perturbations must be small. However, if a regulator studied is only indirectly concerned with the perturbation, it will encounter an effective perturbation that has already been sufficiently buffered by the other regulators, and then even a large or unphysiological perturbation like a gene deletion may be described by a linear theory.

Chapter 6

Properties of optimal expression patterns

In the previous chapter, the model of optimal regulation was presented in its basic form. In this chapter, the model is generalised to changes of the fitness function and to constrained regulation. Furthermore, for limiting cases, the effect of regulation is represented by projection operators, and the model is shown to be invariant against exchanges between the regulators and output variables.

6.1 Perturbation of the fitness

If the fitness $G(x, \alpha, \beta) = F(x, y(x, \alpha), \beta)$ depends on external parameters β, then equation 5.10 becomes

$$\mathrm{d}G_x \quad = \quad = G_{xx}\ \mathrm{d}x + G_{x\alpha}\ \mathrm{d}\alpha + G_{x\beta}\ \mathrm{d}\beta \qquad \text{with} \qquad G_{x\beta} = F_{x\beta} + R_x^{y^T} F_{y\beta} \tag{6.1}$$

The optimal response is still described by equation 5.16, but with

$$\begin{aligned} \mathrm{d}\hat{G}_x &= R_x^{y^T}(F_{yy} R_\alpha^y \mathrm{d}\hat{\alpha} + R_x^{y^T} F_{y\beta} \mathrm{d}\hat{\beta}) + (F_{xy} R_\alpha^y \mathrm{d}\hat{\alpha} + F_{x\beta} \mathrm{d}\hat{\beta}) + (R_{x\alpha}^y \mathrm{d}\hat{\alpha})^T\ F_y \\ &= R_x^{y^T}\, \mathrm{d}\hat{F}_y + \mathrm{d}\hat{F}_x + \mathrm{d}\hat{R}_x^{y^T}\ F_y \end{aligned} \tag{6.2}$$

where

$$\begin{aligned} \mathrm{d}\hat{F}_x &= F_{xy} R_\alpha^y\ \mathrm{d}\hat{\alpha} + F_{x\beta}\ \mathrm{d}\hat{\beta} \\ \mathrm{d}\hat{F}_y &= F_{yy} R_\alpha^y\ \mathrm{d}\hat{\alpha} + F_{y\beta}\ \mathrm{d}\hat{\beta} \end{aligned} \tag{6.3}$$

So changes of the environment are equivalent to changes of the fitness function, as long as they lead to the same changes of the marginal fitness values.

6.2 Constrained regulation

Until here, it was assumed that the relevant output variables y could be regulated independently, that is, any small change dy could be achieved by an appropriate change dx. Now this condition will be relaxed. Instead of solving an optimisation problem with constraints, effective cell variables are introduced, such that the constrained optimisation problem is reduced to an unconstrained problem with external parameters in the fitness function.

The general, constrained problem looks as follows: the regulators x control the output variables z via a function $z = z(x, \alpha)$. The fitness function reads

$$G(x, \alpha, \beta) = F(x, z(x, \alpha), \beta) \tag{6.4}$$

This model implicitly defines an optimal behaviour $\bar{x}(\alpha, \beta) = \text{argmax}_x G(x, \alpha, \beta)$ for a given choice of α and β. At fixed external parameters α, z need not be surjective or injective with respect to x, that is, the control matrix $R^z_x = (\partial z_i(x, \alpha)/\partial x_k)$ may not have full row or column rank. This means that the admissible values of z can be dependent, and different choices of x may yield the same z.

The variables z can always be rewritten, in a region around (x_0, α_0), as $z(x, \alpha) = A(B(x), \alpha)$ such that A is injective and B is surjective (see Figure 6.1): to do so, the response coefficients matrix R^z_x is decomposed into $R^z_x = S\ T$, where S has full column rank and T has full row rank. This can be accomplished, for instance, by singular value decomposition. We can then set

$$\begin{aligned} B(x) &\equiv T \cdot (x - x_0) \\ A(y, \alpha) &\equiv z(T^+ y + x_0, \alpha) \end{aligned} \tag{6.5}$$

where $(\cdot)^+$ denotes the pseudo-inverse. If the fitness H is defined with respect to y as

$$H(x, y, \alpha, \beta) \equiv F(x, A(y, \alpha), \beta)$$

the effective fitness can be rewritten as

$$G(x, \alpha, \beta) = H(x, B(x), \alpha, \beta) \tag{6.6}$$

which has the same form as equation 6.4, but in contrast to $z(x, \alpha)$, $B(x)$ is surjective and independent of external parameters. The former perturbation parameters α now play the role of fitness parameters and could also be lumped with the β to yield a new $\beta^* = \binom{\alpha}{\beta}$. The derivatives of H can be calculated from the derivatives of F and A (see Appendix A.9).

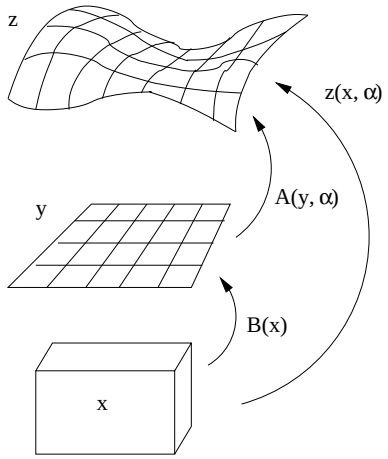


Figure 6.1: An example of constrained regulation. Three regulators (described by a vector $x \in \mathbb{R}^3$) control three output variables z. For fixed environment α, the output state z is located on a two-dimensional surface in $\mathbb{R}^3$. The function $z(x, \alpha)$ is neither surjective nor injective, so the output variables z are dependent, and a change $\mathrm{d}z$ can be achieved by different regulation patterns $\mathrm{d}x$. However, the function $z(x, \alpha)$ can be locally represented by $A(B(x), \alpha)$ such that A is injective (i.e., each small change $\mathrm{d}z$ is represented by exactly one change $\mathrm{d}y$) and B is surjective (each small change $\mathrm{d}y$ can be achieved by a change $\mathrm{d}x$). Now x effectively regulates the new variables y. The fitness $F(x, z(x, \alpha), \beta)$ can be replaced by an effective fitness $H(x, y, \alpha, \beta)$ such that F and H are maximised by the same $\mathrm{d}\bar{x}$ (for details, see text).

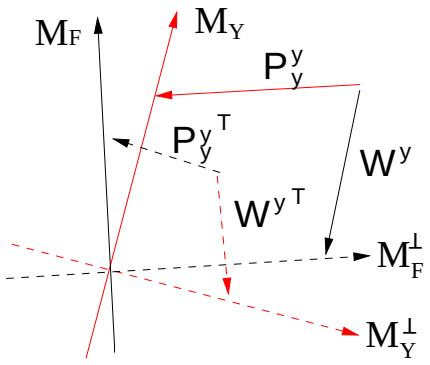


Figure 6.2: Geometrical interpretation of optimal regulation. If the fitness curvatures F_{xx} vanish, the matrices of regulatory coefficients act as projection operators (compare Figure 1.4). The matrix W^y projects a perturbation $\mathrm{d}\hat{y}$ of the output variables to a part $\mathrm{d}\bar{y}$ which cannot be controlled by the regulators. This part is located in $M_F^\perp$, orthogonal to $M_F = \mathrm{span}(F_{yy} R_x^y)$.

6.3 Geometrical interpretation by projections

Metabolic control coefficients are related to projectors in the space of flux distributions (see section 1.4.2). A similar geometrical interpretation also holds for optimal regulation in limiting cases.

1. By optimal regulation, a perturbation $\mathrm{d}\hat{y}$ of the output variables is damped and becomes

$$\mathrm{d}\bar{y} \;=\; W^y\,\mathrm{d}\hat{y} \qquad \text{with} \qquad W^y = 1 - R^y_x\, G^{-1}_{xx}\, R^{y^T}_x F_{yy} \tag{6.7}$$

   The matrix $P^y = 1 - W^y$ maps a perturbation $\mathrm{d}\hat{y}$ to the part which is removed by the regulators. Let us assume that the response coefficients remain constant. Then for soft $F_{xx} \to 0$, P^y converges to a projector on $M_Y = \mathrm{span}(R^y_x)$ while its transposed P^{y^T} projects to $M_F = \mathrm{span}(F_{yy}R^y_x)$ (proof: see Appendix A.8). Accordingly, W^y and its transposed W^{y^T} project to the respective perpendicular spaces $M_Y^\perp$ and $M_F^\perp$. The projection is in general not orthogonal, but reflects the subspaces M_Y and M_F: P^y projects to M_Y, along $M_F^\perp$, while P^{y^T} projects to M_F, along $M_Y^\perp$ (see Figure 6.2). If F_{yy} is isotropic, M_Y and M_F are identical, and the projection is orthogonal.

   How can the projection be interpreted? As $F_{xx} \to 0$, the regulators remove the part of the perturbation $\mathrm{d}\hat{y}$ which is under their control. If R^y_x has full row rank, then $P^y = I$, then the perturbation is removed entirely. If R^y_x has not full row rank, then a part of the perturbation remains.

2. A change $\mathrm{d}\hat{x}$ has the effect $\mathrm{d}\hat{y} = R^y_x \mathrm{d}\hat{x}$. Let us consider a hypothetical external perturbation $\mathrm{d}\hat{\alpha}$ that has the same effect: it would provoke an optimal response $\mathrm{d}\bar{x}$ which can be seen as a projection of $-\mathrm{d}\hat{x}$. We first assume that the regulators compensate entirely for the perturbation (as in section 5.1.3). According to equation 5.23, the optimal way to remove $\mathrm{d}\hat{y}$ is

$$\mathrm{d}\bar{x} = -P^x_x \mathrm{d}\hat{x} \quad \text{with} \quad P^x_x = F^{-1}_{xx} R^{y^T}_x (R^y_x F^{-1}_{xx} R^{y^T}_x)^{-1} R^y_x \tag{6.8}$$

   Irrespective of the choice of F_{xx}, the matrix P^x_x is a projector to $M_F = \mathrm{span}(F^{-1}_{xx} R^{y^T}_x)$, while $P^{x^T}_x$ projects to $M_R = \mathrm{span}(R^{y^T}_x)$. On the other hand, the optimal response (to maximise the total fitness, as in section 5.1.2) reads

$$\mathrm{d}\bar{x} = -G^{-1}_{xx} R^{y^T}_x F_{yy} R^y_x \mathrm{d}\hat{x} = -Q^x_x \mathrm{d}\hat{x} \tag{6.9}$$

   By inserting $G_{xx} = (F_{xx} + R^{y^T}_x F_{yy} R^y_x)$, we see that for small $F_{xx} \to 0$, Q^x_x becomes an orthogonal projector to the eigenvectors of $R^{y^T}_x F_{yy} R^y_x$ with non-vanishing eigenvalues.

6.4 Invariance against reassignment of regulators

To calculate optimal regulation patterns, an artificial distinction has been made between regulatory variables, which follow optimality, and output variables, which depend on the regulators and on the environment. There is no such distinction in reality, so the theory must at least be invariant to a redistribution of regulators and system variables, together with an appropriate new choice of the fitness function: if a regulator becomes an output variable, it is supposed to keep its optimal behaviour, and if an output variable becomes a regulator, its behaviour must be optimal with respect to the new fitness function. Above, the fitness function was assumed to be decomposable into $F(x, y) = U(x) + V(y)$, and this property must also be maintained.

The model indeed fulfils the invariance postulate, with a slight restriction: let us assume that the fitness contributions U and V can be further split into sums

$$\begin{aligned} U(x) &= U_1(x^{(1)}) + U_2(x^{(2)}) + ... \\ V(y) &= V_1(y^{(1)}) + V_2(y^{(2)}) + ... \end{aligned} \tag{6.10}$$

where each term depends on minimal subsets $x^{(i)}$ of regulators or minimal subsets $y^{(k)}$ of variables. These subsets are called fitness-closed. The model is invariant against reassignments of regulators and variables as long as no fitness-closed set is split into both regulators and variables.

For the prove, it must be shown that (1) regulators z forming a fitness-closed set can become output variables and (2) output variables z forming a fitness-closed set can become regulators, while, for an appropriate fitness function, the optimality postulate still leads to the same system behaviour.

1. The fitness can be decomposed according to

$$\begin{aligned} G\left(\binom{x}{z}, \alpha\right) &= F\left(\binom{x}{z}, y\left(\binom{x}{z}, \alpha\right)\right) \\ &= U^{(x)}(x) + U^{(z)}(z) + V\left(y\left(\binom{x}{z}, \alpha\right)\right) \end{aligned} \tag{6.11}$$

   if the regulators z form a fitness-closed set. For any given α, they assume their optimal values $\bar{z}(\alpha)$, so the other regulators x have to maximise

$$\begin{aligned} G\left(\binom{x}{\bar{z}(\alpha)}, \alpha\right) &= U^{(x)}(x) + U^{(z)}(\bar{z}(\alpha)) + V\left(y\left(\binom{x}{\bar{z}(\alpha)}, \alpha\right)\right) \\ &= U^{(x)}(x) + V^*(y^*(x, \alpha)) \end{aligned} \tag{6.12}$$

where the new output variables read

$$y^* = \begin{pmatrix} y\left(\binom{x}{\bar{z}(\alpha)}, \alpha\right) \\ \bar{z}(\alpha) \end{pmatrix} \tag{6.13}$$

2. The fitness reads

$$G(x,\alpha) \;=\; F\left(x, \begin{pmatrix} y(x,\alpha) \\ z(x,\alpha) \end{pmatrix}\right) = U(x) + V\left(\begin{pmatrix} y(x,\alpha) \\ z(x,\alpha) \end{pmatrix}\right) \tag{6.14}$$

If the regulators $x = \bar{x}(\alpha)$ behave optimally, the variables z assume the values $\bar{z}(\alpha) = z(\bar{x}, \alpha)$. A term $\tilde{U}(z) = |z - \bar{z}(\alpha)|^2$ is added, and the new fitness function

$$G^*\left(\begin{pmatrix} x \\ z \end{pmatrix}, \alpha\right) = U(x) + \tilde{U}(z) + V\left(\begin{pmatrix} y(x,\alpha) \\ \bar{z}(\alpha) \end{pmatrix}\right) = U^*\left(\begin{pmatrix} x \\ z \end{pmatrix}\right) + V^*(y(x,\alpha)) \tag{6.15}$$

is maximal if the old $\bar{x}(\alpha)$ maximise G and if the new regulators z behave like the former output variables $z(x,\alpha)$.

Chapter 7

Time-dependent expression

The model presented so far described the regulation of stationary states, while the expression data studied before consisted of time series. In this chapter, it is shown that the model, with slight modifications, also applies to time-dependent perturbations. To account for the time-dependent behaviour of a metabolic system to be regulated, the response coefficients are replaced by frequency-dependent response coefficients which describe how cell variables respond to oscillatory parameter changes. Summation and connectivity theorems are derived for the frequency-dependent control coefficients.

7.1 Time-dependent gene expression

The stress response data Gasch et al. [32] contain expression time series after different changes of external conditions. For most of these experiments, the loadings of the first two principal components can be fitted by superpositions of two exponentially relaxing curves (see Figure 7.1). The time scales of the faster exponential relaxation (see Table 7.1) are in the order of 10 minutes, suggesting that gene expression limits the velocity of the shock response. Time series behind global gene expression (derived from expression data by clustering or singular value decomposition) have been explained by linear or nonlinear dynamical systems [20] [114] [113] [48] [106]. In contrast to a dynamical modelling of gene expression, this chapter studies how the concept of optimal regulation can be generalised to time series.

The model in chapter 5 assumed small stationary perturbations around an optimal steady state. In reality, perturbations will occur in time and will lead to a temporary response. The perturbations may represent coordinated processes in the cell, e.g., changes during the cell cycle, or stochastic fluctuations of the environment, which can be characterised by a frequency spectrum. If the perturbations are slow, the regulators may respond to them

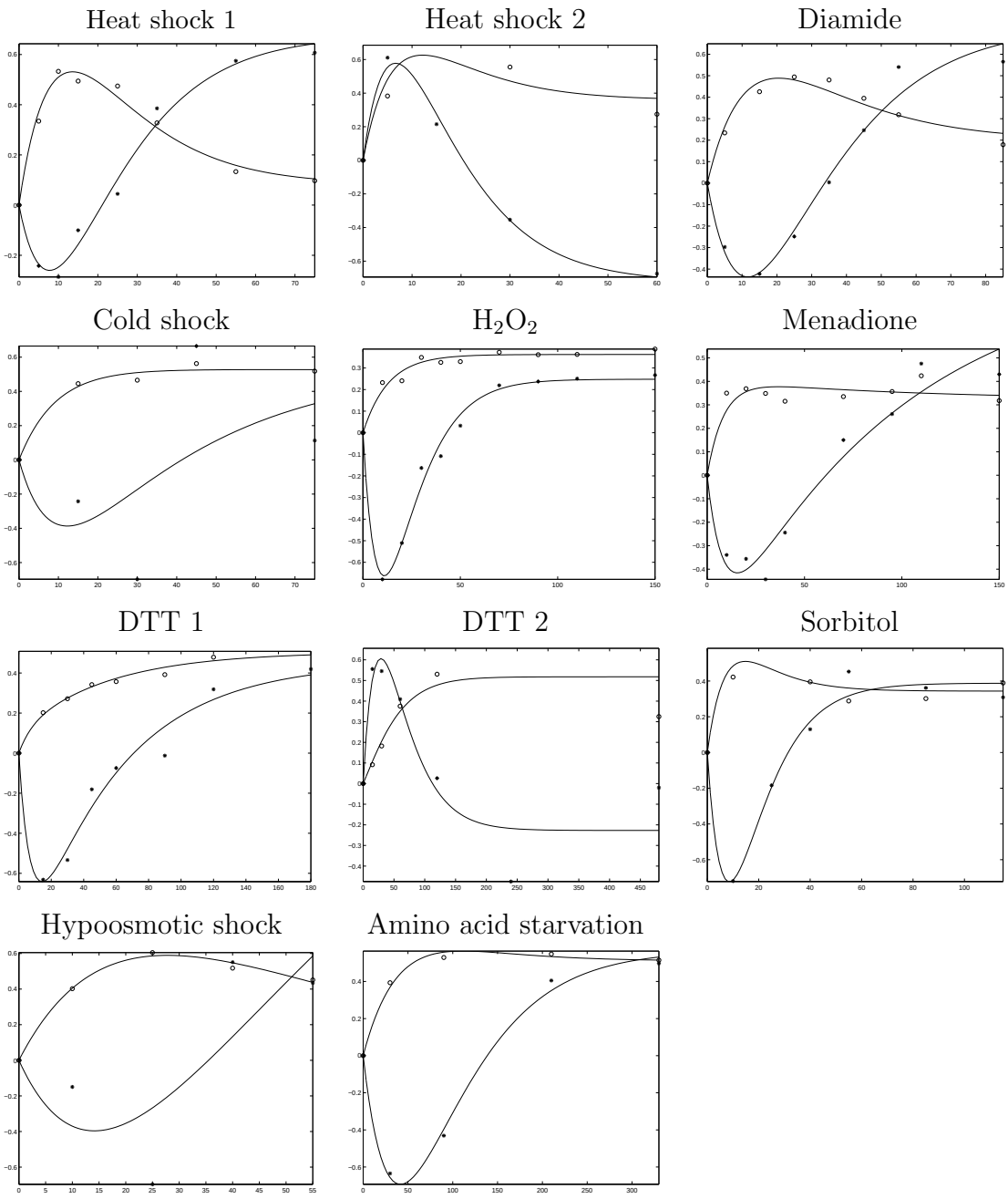


Figure 7.1: Expression modes in the environmental changes data [32]. Each diagram shows the time series (abscissa: time in minutes) of the first two PCA modes (stars and circles) from one of 11 stress experiments (compare Table 7.1). The time series can be fitted by linear superpositions of two exponentially decreasing functions: the fits are shown by solid lines.

	Time series	$\tau_1(min)$	$\tau_2(min)$
1	Heat shock 1	11	14
2	Heat shock 2	7.5	11
3	Diamide	17	17
4	Cold shock	9.1	37
5	H_2O_2	11	14
6	Menadione	8.4	96
7	DTT 1	7.4	55
8	DTT 2	30	35
9	Sorbitol	8.3	12
10	Hypoosmotic shock	18	60
11	Amino acid starvation	48	57
12	Diauxic shift	$3.8 \cdot 10^7$	$1.0 \cdot 10^8$
13	Nitrogen depletion	140	170
14	stationary phase 1	270	$1.0 \cdot 10^4$
15	stationary phase 2	980	3000

Table 7.1: Time constants for exponential modes behind stress response data Gasch et al. [32] (compare Figure 7.1). In the last five experiments, the time scale relevant for expression ($< 1/2$ hour) was not resolved by the experiments. In the experiments 1-10, the small time constant is between 7.4 and 11 minutes in 7 out of 10 cases, probably representing the time scale of gene expression.

in a quasi-stationary way. For faster perturbations, the time-delay of regulation must be considered: the usual turnover times of mRNA and also the time scale of translation are between minutes and tens of minutes, while the effects of enzyme changes propagate through the metabolic network within seconds or minutes. The regulators may be optimised for fluctuating perturbations of certain frequencies, as they typically occur in the evolutionary environment: for slow perturbations, the response should resemble the one for stationary states, while for fast fluctuations, it should tend to anticipate the future course of the perturbation. For cell-cycle-related processes, this is possible because a central cell-cycle clock can activate genes already before their products are needed. To cope with fast unpredictable perturbations, the further course of the perturbation can be anticipated by an overshooting response.

7.2 Frequency-dependent control coefficients

The response coefficients defined in metabolic control theory describe asymptotic effects of a parameter change, for instance, the increased expression of an enzyme. In reality, the effects need some time to propagate through the network. If the metabolic system is perturbed by small oscillatory parameter changes, its response can be characterised, in a linear approximation, by frequency-dependent metabolic response coefficients, which will

be defined here. They describe perturbations of a non-oscillatory state and thus must not be confused with response coefficients describing the characteristics of limit cycles [96]. The present approach is closely related to [98], where the transmission of an oscillatory influx through small enzyme chains was studied.

Let us consider a system of differential equations characterised by time-independent parameters p

$$\dot{y}(t) = f(y(t), p) \tag{7.1}$$

With a standard parameter set p_0, the parameters can be expressed by $p = p_0 + \tilde{p}$. Let us assume that for parameters $p_0 + \tilde{p}$, there is a stationary state

$$x(\tilde{p}) \tag{7.2}$$

fulfilling $0 = f(x(\tilde{p}), p_0 + \tilde{p})$. The response coefficients are defined as

$$(R_p^x)_n^m \equiv \frac{\partial x_m}{\partial \tilde{p}_n}|_{\tilde{p}=0} \tag{7.3}$$

Now let us assume small time-dependent perturbations of a stable steady state x_0 at parameters p_0. As the problem is invariant against a shift of time, only perturbations $\tilde{p}e^{i\omega t}$ (eigenvectors of the time-shift operator) with complex coefficients (elements of a vector $\tilde{p}$) need to be considered, while other perturbations can be superposed from them by Fourier synthesis. For parameters $p(t) = p_0 + \tilde{p}e^{i\omega t}$, the solution $x(t)$ of the differential equation system

$$\dot{x}(t) = f(x(t), p(t)) \tag{7.4}$$

can be written as $x(t) = x_0 + \tilde{x}e^{i\omega t}$ plus higher order terms, which can be neglected for small perturbations $\tilde{p}$. In analogy to (7.2), we get a function $\tilde{x}(\tilde{p})$ which can be used to define complex, frequency-dependent response coefficients

$$(R_p^x)_m^n(\omega) \equiv \frac{\partial \tilde{x}_n(\omega)}{\partial \tilde{p}_m} \tag{7.5}$$

For being more specific, let us consider the frequency-dependent control coefficients for a metabolic system

$$\dot{s} = Nv(s, p) \tag{7.6}$$

with elasticity matrices $\epsilon = \nabla_s^T v$ and $\pi = \nabla_p^T v$. The stationary state S at parameters p_0 fulfils

$$0 = Nv(S, p_0) \tag{7.7}$$

A perturbation $p(t) = p_0 + \tilde{p}\ e^{i\omega t}$ is applied, and the resulting $v(t)$ is split into

$$v(t) = v(S, p_0) + \Delta v(t) \tag{7.8}$$

The ansatz

$$s(t) = S + \tilde{S}e^{i\omega t} \tag{7.9}$$

with a complex amplitude $\tilde{S}$ yields, for small perturbations,

$$\Delta v(t) \approx (\epsilon\tilde{S} + \pi\tilde{p})e^{i\omega t} = \tilde{v}e^{i\omega t} \tag{7.10}$$

so

$$\dot{s} \;=\; i\omega\ \tilde{S}\ e^{i\omega t} = N(v(S, p_0) + \tilde{v}) \approx N(\epsilon\tilde{S} + \pi\tilde{p})e^{i\omega t} \tag{7.11}$$

Cancelling down $e^{i\omega t}$ on both sides yields, in first order

$$\tilde{S} \;=\; -(N\epsilon - i\omega)^{-1}N\pi\tilde{p} \tag{7.12}$$

$$\rightarrow \partial\tilde{S}/\partial\tilde{p} \;=\; -(N\epsilon - i\omega)^{-1}N\pi \tag{7.13}$$

Here the symbolic notation $\partial\tilde{S}/\partial\tilde{p}$ denotes a matrix $\nabla^T_{\tilde{p}}\tilde{S}$. If the metabolite concentrations are constrained by conservation relations, then a similar calculation shows that

$$\partial\tilde{S}/\partial\tilde{p} = -L(M^0 - i\omega)^{-1}N^0\pi \tag{7.14}$$

where $M^0 = N^0\epsilon L$. Considering parameter changes $\tilde{p}$ that act specifically on single reactions, we can define frequency-dependent control coefficients.

$$C^S(\omega) \;\equiv\; \frac{\partial\tilde{S}/\partial\tilde{p}}{\partial\tilde{v}/\partial\tilde{p}} = -L(M^0 - i\omega)^{-1}N^0 \tag{7.15}$$

$$C^J(\omega) \;\equiv\; \frac{\mathrm{d}\tilde{v}/\mathrm{d}\tilde{p}}{\partial\tilde{v}/\partial\tilde{p}} = 1 + \epsilon C^S(\omega) \tag{7.16}$$

These equations resemble the known formulae 1.14 for control coefficients, but here the Jacobian $N\epsilon$ is modified by a term $-i\omega$, which vanishes again in the stationary case $\omega \rightarrow 0$. The coefficients fulfil the summation and connectivity theorems

$$C^S(\omega)K \;=\; 0 \tag{7.17}$$

$$C^J(\omega)K \;=\; K \tag{7.18}$$

$$C^S(\omega)(\epsilon L - i\omega N^{0^+}) \;=\; -L \tag{7.19}$$

$$C^J(\omega)(\epsilon L - i\omega N^{0^+}) \;=\; -i\omega N^{0^+} \tag{7.20}$$

where $N^{0^+} \equiv N^{0^T}(N^0\ N^{0^T})^{-1}$. The product $N^0\ N^{0^T}$ is invertible because N^0 has full row rank by definition.

Frequency-dependent response coefficients with respect to $\tilde{p}$ are defined similarly by

$$R_p^S(\omega) \equiv \frac{\partial \tilde{S}}{\partial \tilde{p}} = C^S(\omega)\ \pi(\omega) \tag{7.21}$$

$$R_p^J(\omega) \equiv \frac{\partial \tilde{S}}{\partial \tilde{p}} = C^J(\omega)\ \pi(\omega) \tag{7.22}$$

7.3 Optimal time-dependent regulation

The knowledge about frequency-dependent control can be used to calculate the optimal response to time-dependent perturbations of optimal stationary states. The perturbation $\Delta\alpha(t)$ is expanded into oscillatory modes $\Delta\alpha(t) = \Delta\alpha_\omega e^{i\omega t}$, and $\Delta x(t)$ and $\Delta y(t)$ are expanded likewise. For given time-dependent perturbations $\Delta\alpha(t)$ and responses $\Delta x(t)$, the time-averaged fitness change

$$\langle \Delta F \rangle = \frac{1}{2}\langle \Delta x(t)^T F_{xx} \Delta x(t) + \Delta y(t)^T F_{yy} \Delta y(t) \rangle \tag{7.23}$$

can be expressed by the frequency-dependent amplitudes Δx_ω and Δy_ω

$$\langle \Delta F \rangle = \frac{1}{4\pi} \int_{-\infty}^{\infty} \Delta x_\omega^\dagger F_{xx} \Delta x_\omega + \Delta y_\omega^\dagger F_{yy} \Delta y_\omega \ \mathrm{d}\omega \tag{7.24}$$

The symbol † indicates the adjoint (i.e., the transposed and complex conjugate) matrix. With the frequency-dependent response coefficients, Δy can be expanded in first order

$$\Delta y_\omega = R_\alpha^y(\omega)\Delta\alpha_\omega + R_x^y(\omega)\Delta x_\omega \tag{7.25}$$

By inserting Δy_ω, the integrand in equation (7.24) becomes

$$\Delta x_\omega^\dagger G_{xx}(\omega)\Delta x_\omega + \Delta x_\omega^\dagger G_{x\alpha}(\omega)\Delta\alpha_\omega + \Delta\alpha_\omega^\dagger G_{\alpha x}(\omega)\Delta x_\omega + \Delta\alpha_\omega^\dagger G_{\alpha\alpha}(\omega)\Delta\alpha_\omega \tag{7.26}$$

with

$$\begin{aligned} G_{xx}(\omega) &= F_{xx} + R_x^{y^\dagger}(\omega) F_{yy} R_x^y(\omega) \\ G_{x\alpha}(\omega) &= R_x^{y^\dagger}(\omega) F_{yy} R_\alpha^y(\omega) \\ G_{\alpha x}(\omega) &= G_{x\alpha}^\dagger(\omega) \\ G_{\alpha\alpha}(\omega) &= R_\alpha^{y^\dagger}(\omega) F_{yy} R_\alpha^y(\omega) \end{aligned} \tag{7.27}$$

The optimal response can be determined for each frequency ω separately

$$\Delta\bar{x}_\omega = -G_{xx}(\omega)^{-1} G_{x\alpha}(\omega)\ \Delta\alpha_\omega = W(\omega)\ \Delta\alpha_\omega \tag{7.28}$$

This resembles the result (5.12) for the stationary case, except for the frequency-dependent response coefficients. The optimal response in time can be expressed by a convolution integral

$$\Delta\bar{x}(t) = \int_{-\infty}^{\infty} K(t-t')\ \Delta\alpha(t')\ \mathrm{d}t' \tag{7.29}$$

with a matrix-valued kernel $K(\tau)$ determined by Fourier transformation

$$K(\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega\tau}\ W(\omega)\ \mathrm{d}\omega \tag{7.30}$$

The optimal behaviour described here is not constrained to be causal, i.e., to start after the perturbation. Quite contrarily, optimality makes it necessary to anticipate the perturbations: for cell processes that follow a prescribed choreography, such as the cell cycle, this can be biologically realised. On the other hand, the cell can also cope optimally with unpredictable perturbations, but this problem is quite involved: mathematically, it requires a constrained optimisation with respect to an ensemble of time-dependent perturbations.

Another interesting question, which has not been treated here, is the compromise between regulators acting on different timescales and with different cost functions. For instance, metabolic adaptations can be effected fast, by allosteric control of enzymes, or slowly, by expression changes of enzymes that adapt the operating point of the metabolic system. The total energetic effort that is put into both systems may be minimised by distributing the control between them in an optimal way.

Chapter 8

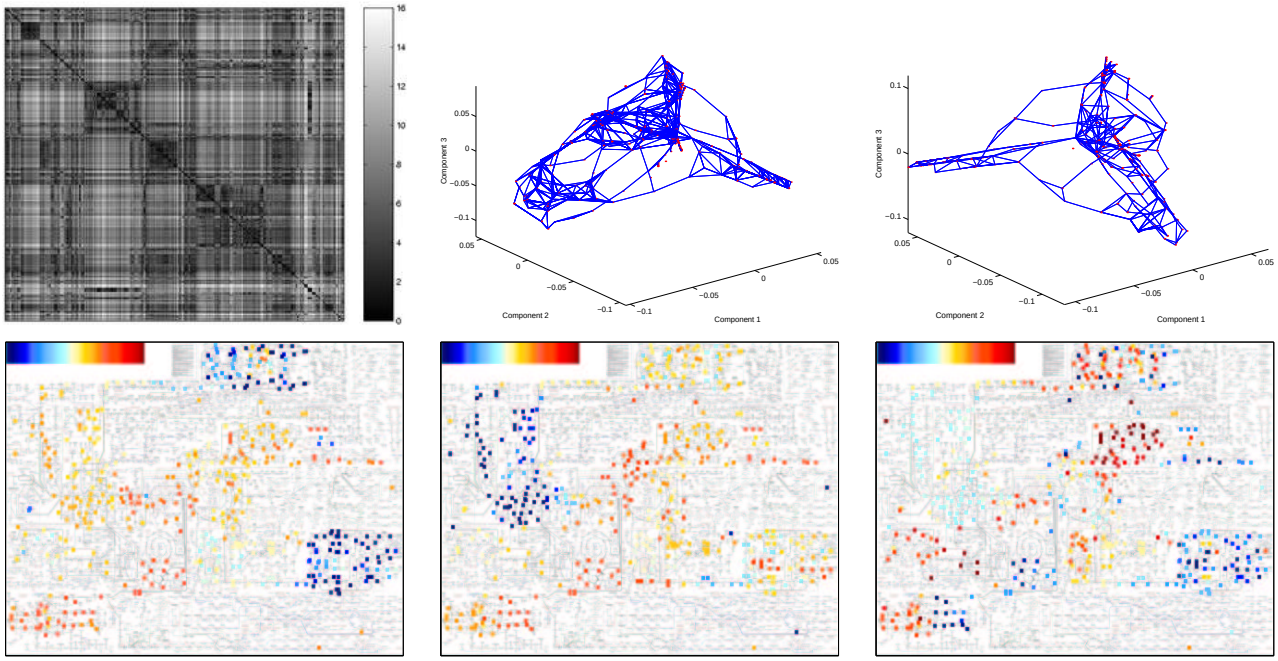
Calculation of control coefficients

A relation between optimal expression profiles and the response coefficients will be claimed in chapter 9, and large matrices of response coefficients will be necessary to test it. In this chapter, metabolic control coefficients are calculated for a large metabolic network, and their distributions and statistical dependencies are studied. I shall refer to them as "simulated control coefficients" because very simple assumptions are made about the elasticities. The control coefficients are almost sparse and reflect the structure of the metabolic network. Coefficients within a small subnetwork depend only weakly on modelling of the surrounding network. The frequency-dependent control coefficients show frequency-dependent phase-shifts, but only weak resonance.

8.1 Metabolic control coefficients

What are the statistical properties of metabolic control coefficients? According to the theorems 1.16, metabolic control coefficients reflect the network structure, the stoichiometry and reversibility of chemical reactions, and the linearised reaction kinetics. General properties of the network topology may force some of the control coefficients to vanish, for instance, those between unconnected subnetworks. The control coefficients are statistically dependent: the theorems 1.16 induce linear correlations between them. Textbooks on cell biology (see, e.g., [108]) sometimes put forward the notion of "rate-limiting steps" in metabolic pathways, which implicitly claims that control coefficients are almost sparse. This sparsity assumption will be tested below.

I calculated metabolic control coefficients for a large network to study their statistics and to relate them to expression data. According to equation 1.14, control coefficients are determined by the stoichiometric matrix N, describing the network topology, and by



Component 1 Component 2 Component 3

Figure 8.1: Topology of a reaction network describing metabolism in yeast. Top left: The network defines distances D_{ik} between reactions by the length of shortest paths. The network distance matrix is shown. Top centre: Graph of chemical reactions. Reactions which share a metabolite are connected by an edge. Positions of the graph nodes were determined by multidimensional scaling: the reactions were mapped to points in $\mathbb{R}^3$ such that the network distances D_{ik} are approximately represented by the Euclidean distances between the nodes. Each reaction is represented by 3 coordinates. Top right: Graph of metabolites for the same metabolic network. The nodes and edges of this graph correspond to metabolites and reactions, respectively. Bottom: The Boehringer chart [83] is optimised to represent reactions in two dimensions, but some metabolites appear several times, and distances in the map need not reflect the true network distances. Nevertheless, the network considered here is well recognisable on the map. Each diagram shows one of the 3 coordinates (colour-coded) from multidimensional scaling: regions of the network appear as point clouds of similar colour.

the reaction elasticities ϵ, describing the linearised kinetics. To build a network describing central parts of the yeast metabolism, 566 reactions were chosen from the LIGAND database [64][1]. The nominal directions of the 289 reversible reactions may differ from the flux directions under physiological conditions. The stoichiometric matrix defines a graph between reactions, in which reactions sharing substrates or products are connected by an edge. The topology of this network and the corresponding network distances D_{ik}, defined by the shortest path lengths on the graph, are shown in Figure 8.1.

8.1.1 Choice of the elasticities

For almost all of the reactions considered, the kinetic parameters are unknown, so the reaction elasticities had to be guessed. To do so, it was assumed that the velocity of a reaction depends only on the participating metabolites, while other regulatory influences were neglected. For irreversible reactions, the product has no influence on the reaction velocity. For simplicity, the elasticities between a reaction and its substrates and products were set to 1 and -1, respectively, while for irreversible reactions, the value -1 was replaced by -0.001. If a value of zero had been used, the elasticity matrix ϵ might not have its full rank, and the Jacobian matrix M^0 in equation 1.14 could not be inverted.

Setting the elasticities to ± 1 is only a rough guess. To study how different choices of the elasticities would influence the results, random values for the elasticities were drawn from a log-normal distribution. By this Monte Carlo simulation, it can be tested how the statistical properties of control coefficients depend on the particular choice of the elasticities. In the random assignments, the signs from the basic "deterministic" model were kept and only the absolute values were varied: to do so, the elements were multiplied by independent log-normal random numbers $\exp(\eta)$ where η is normally distributed with zero mean and a standard deviation $\sigma_\epsilon = 1$.

Is it plausible to assume log-normal elasticities? Figure 8.2 shows that the assumption approximately holds for the elasticities from the glycolysis model of Hynne et al. [52], if positive and negative values are considered separately. To test which properties of the control coefficients can be predicted from based on random elasticities, the same model was used: Figure 8.3 shows the control coefficients for the glycolysis model along with simulations based on random elasticities, sampled from 100 simulation runs. Signs that had been consistently found in more than 55 runs were used for predicting the true signs, and in almost all cases, these prediction were correct. The prediction of the absolute values, by the geometric mean over 100 simulation runs, was less reliable. The reason

[1] EC numbers for which no yeast ORF is annotated were omitted. The network contains 447 metabolites, 101 of which were considered external because they participate in only one reaction. No cell compartments were considered.

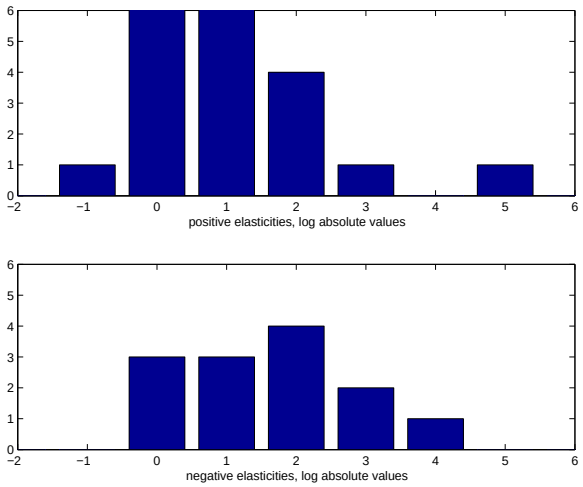


Figure 8.2: Reaction elasticities in the yeast glycolysis model from [52] are approximately log-normal. Hynne et al. determined kinetic parameters for reactions within and around glycolysis: the model structure is shown in Figure 5.7. The diagrams show histograms of the $\log_{10}$ absolute non-vanishing elasticities with positive (top) and negative signs (bottom): the distributions are close to normal with means and standard deviations 1.45 ± 1.27 (positive values) and 2.25 ± 1.19 (negative values).

is that single large elasticities lead to strong rows or columns in the control coefficients matrices. Normalising the rows and columns of the control coefficient matrices by their standard deviations improved the predictions. In this example, knowledge about the network topology gave hints about dependencies between control coefficients, but for predicting their absolute values, the true elasticities had to be known.

8.2 Distributions and correlations of control coefficients

The control coefficients for the metabolic network in yeast were sampled from ten simulation runs. Figure 8.4 shows the averaged control coefficients on Pyruvate: for averaging, the most frequent signs and geometric means of the absolute values were chosen from the simulation runs.. The control coefficient matrix C^J is shown in Figure 8.5. The histograms indicate that the control coefficients have almost sparse distributions.

Due to the theorems 1.16, the matrices C^S and C^J do not have their full rank, so the control coefficients are correlated. The correlations can be represented by splitting the flux control coefficients matrix C^J into a product

$$C^J = K\ Q \tag{8.1}$$

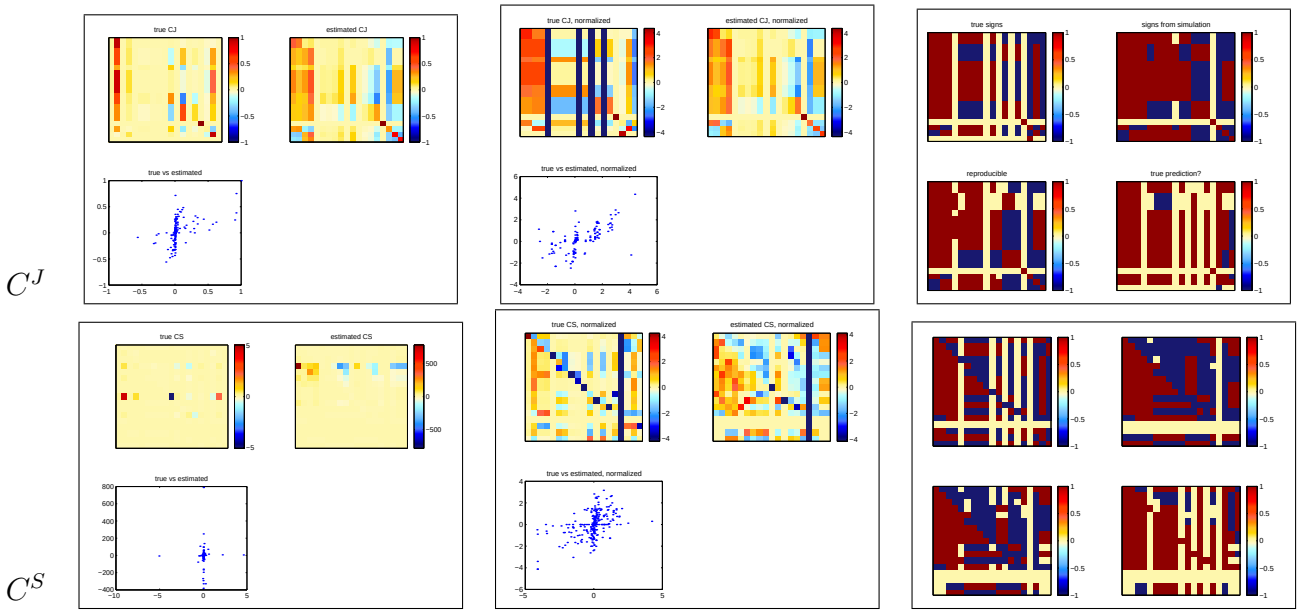


Figure 8.3: Simulation of control coefficients for the glycolysis model [52] (compare Figures 5.7 and 8.2). The control coefficients for the internal metabolites and their reactions are compared to "simulated" coefficients, calculated from randomly chosen elasticity values. Top left: True and simulated flux control coefficients (simulations were averaged over 100 runs). The scatter-plot, below, shows only a weak correlation between them. Top centre: If the matrix rows and columns are normalised by their standard deviation, the prediction by the simulations becomes better. Top right: Signs of flux control coefficients (brown, beige, and blue indicate the values +1, 0, and -1, respectively). The upper two boxes show the true signs and average simulated signs. The lower two boxes show the simulated signs that were consistent for more than 55 simulations, and the prediction success (brown and blue indicate true and false predictions). Bottom row: Concentration control coefficients. The diagrams are similar to the ones on top.

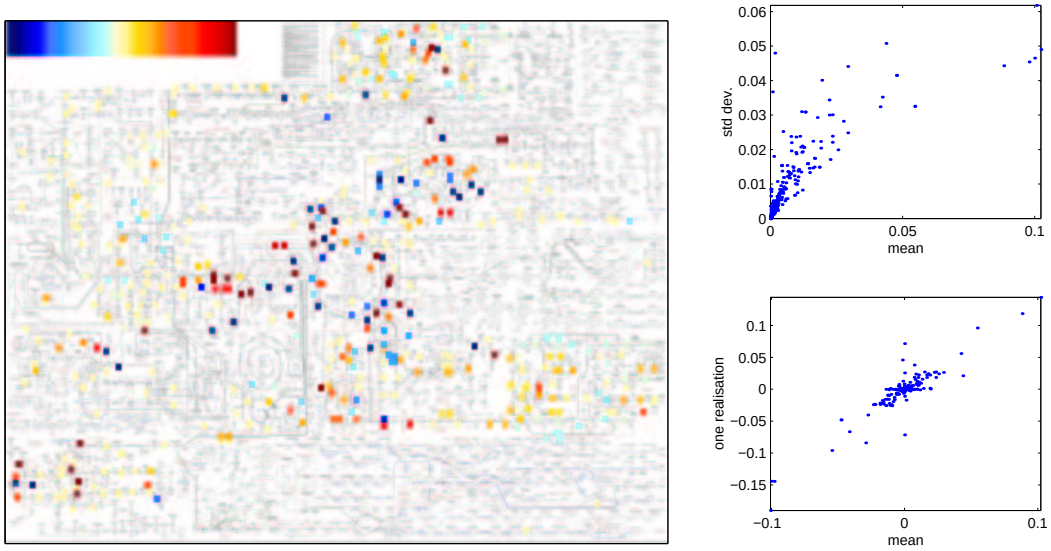


Figure 8.4: Monte Carlo simulation of control coefficients. Concentration control coefficients on Pyruvate were calculated for ten random assignments of the reaction elasticities. Left: Mean control coefficients on Pyruvate. The colour scale refers to arsinh-transformed values. Right: Control coefficients on Pyruvate. Top: The control coefficients were calculated from 10 random assignments. Their mean values (abscissa) and standard deviations (ordinate) are in the same order of magnitude. Bottom: Control coefficients calculated with fixed elasticities ± 1 are plotted versus the mean values calculated from the ten random assignments of the elasticities.

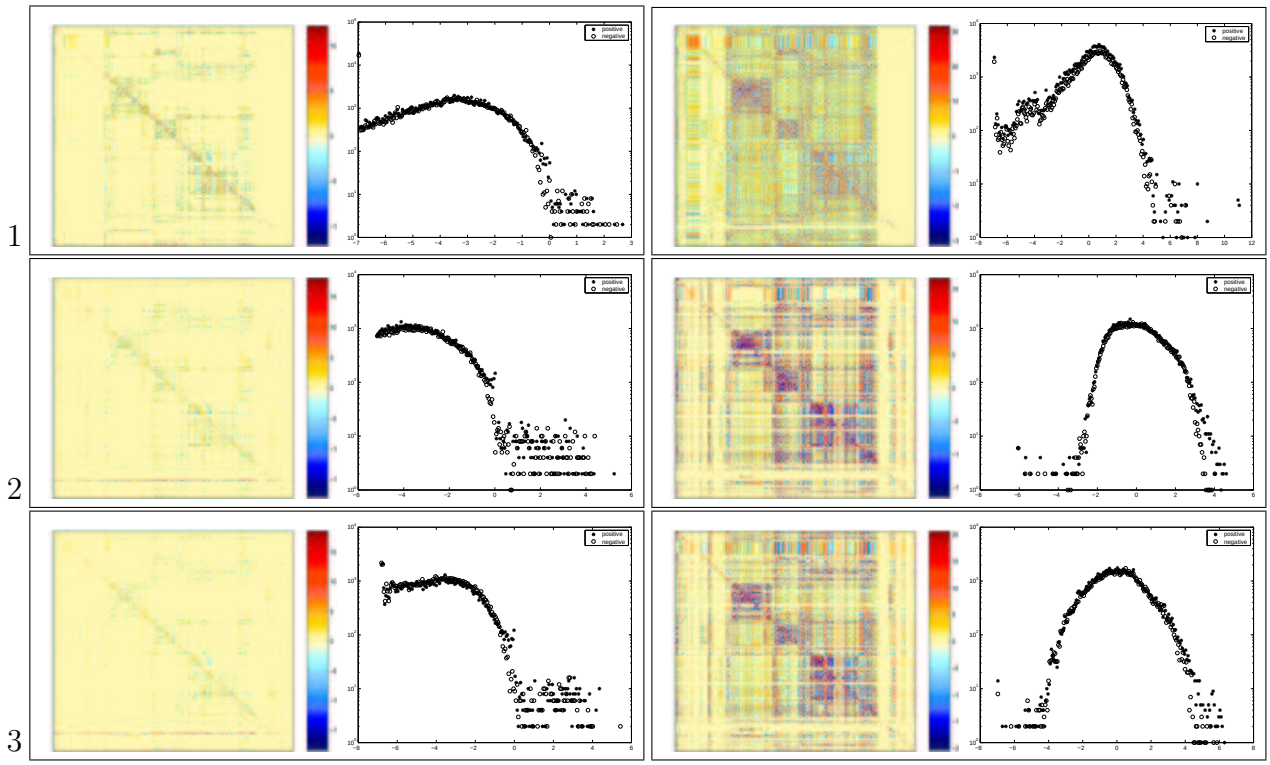


Figure 8.5: Flux control coefficient matrix C^J for the yeast metabolic network. The diagrams show the matrix along with histograms of the log absolute values of non-vanishing elements. Left column: The boxes correspond to (1) the "deterministic" model with elasticity values ± 1. (2) one random assignment of the elasticities (3) average over 10 random assignments. For averaging, the geometric mean of the absolute values and the most frequent sign were chosen for each element. The colour scale refers to arsinh-transformed elements of C^J. In the histograms (log-scale), positive and negative elements are shown by crosses and circles, respectively. Right boxes: Here each row of C^J was scaled by the geometric mean of its absolute values.

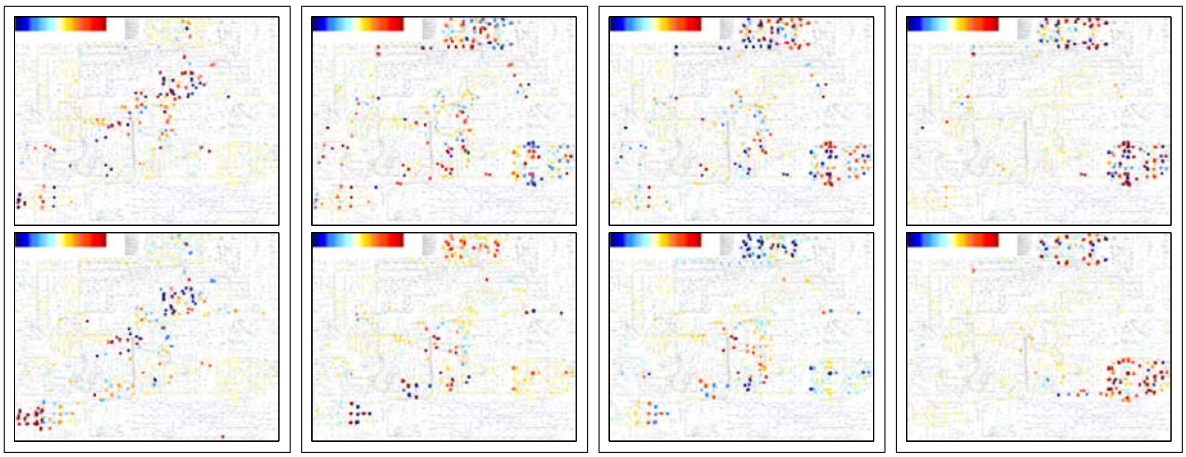


Figure 8.6: Uncorrelated flux modes k_i and control profiles q_i, determined from C^J by a singular value decomposition. The control coefficients matrix is decomposed into $C^J = KQ$ where the columns k_i of K ("uncorrelated flux modes") are orthogonal on each other, that is, linearly uncorrelated, and the same holds for the rows q_i of Q ("uncorrelated control profiles"). The boxes show the first four uncorrelated flux modes (top) and the corresponding control profiles (bottom).

where K and Q have full column and row rank, respectively. By a singular value decomposition of C^J, the kernel matrix K can be chosen to have orthogonal, and thus linearly uncorrelated columns k_i, which I shall call "uncorrelated flux modes". At the same time, also the rows q_i of Q (which will be called "uncorrelated control profiles") are linearly uncorrelated. What is the meaning of this decomposition? A flux perturbation $\mathrm{d}v$ by a parameter change yields a stationary flux change

$$\mathrm{d}J = C^J \mathrm{d}v = KQ\mathrm{d}v = \sum_i k_i (q_i \mathrm{d}v) \tag{8.2}$$

The remaining flux change $\mathrm{d}J$ is a superposition of the uncorrelated flux modes, while for each of them, the corresponding control profile determines how strong it will respond to particular perturbations $\mathrm{d}v$. The first uncorrelated flux modes and control profiles are shown in Figure 8.6: they represent large regions of the metabolic network.

Both the control coefficients and their correlations reflect the network topology. In Figure 8.7, top centre and right, absolute control coefficients $|(C^J)^i_k|$ are plotted versus the network distance D_{ik} between the reactions. On average, the control decreases with higher distances, and the same holds for the correlated control on other reactions or metabolites (shown on bottom).

Modelling of isolated metabolic subsystems is based on the tacit assumption that the control coefficients depend weakly on distant parts of the network. However, it is not obvious whether an incomplete model network can yield realistic control coefficients at

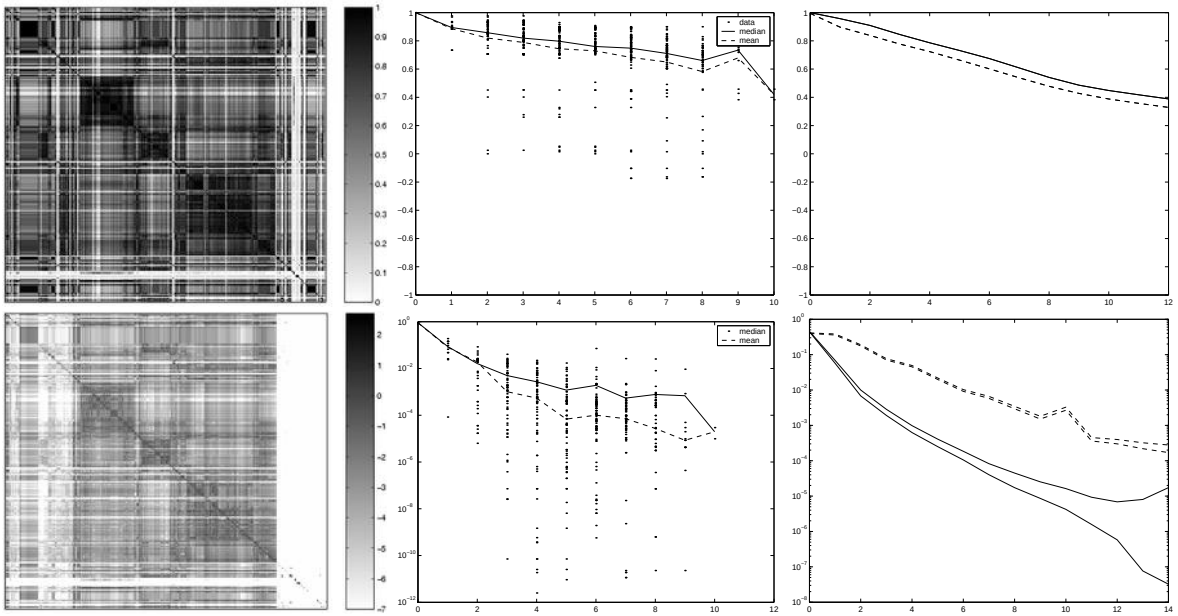


Figure 8.7: Flux control coefficients and their correlations reflect the network topology. High values in the control profiles (rows of C^J) tend to be localised in the network. Top left: The matrix of absolute flux control coefficients $|C^J|$ (logarithmic colour-scale) resembles, by its structure, the network distance matrix (shown in Figure 8.1). Top centre: Control on the reaction Oxaloacetate $\leftrightarrow$ Pyruvate. The absolute control coefficients $|(C^J)^i_k|$ are plotted against the network distance D_{ik} between the reactions. Both the median (solid line) and the mean (dashed line) decrease with higher distances. Top right: The same, for all elements from C^J. Bottom: Correlation between control coefficients reflect the network topology. Left: Absolute correlations between flux control coefficients (columns of C^J). Like the absolute control coefficients themselves (compare Figure 8.7), the correlations between them (show on the ordinate, in log. scale) decrease with the network distance. The diagrams refer to correlated control on the reaction Oxaloacetate $\leftrightarrow$ Pyruvate (centre) and on all reactions (right). For the latter diagram, median and mean values were calculated with (lower curves) and without (upper curves) the vanishing elements.

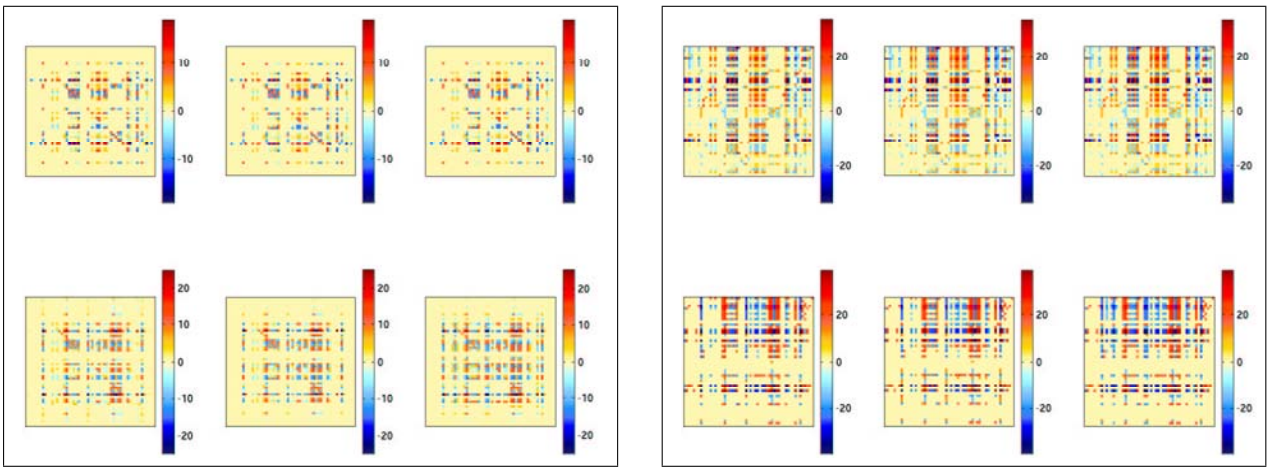


Figure 8.8: Control coefficients for a metabolic subnetwork. Control coefficients were calculated from metabolic networks containing 50, 52, 57, 63, 86, and 144 reactions around the metabolite Pyruvate, each network being a subnetwork of the following one. The diagrams show the matrices of flux (left) and concentration (right) control coefficients for the metabolites and reactions in the smallest network. The colour-scheme is the same as in Figure 8.5. For networks containing 63 or more reactions, the control coefficients remain almost constant.

all. As a test, control coefficients were calculated for networks of different size, each being a subnetwork of the following one. As an example, Figure 8.8 shows networks around the metabolite Pyruvate. The networks were determined by first choosing the n metabolites whose C^S correlate best to those of Pyruvate, and choosing then the largest connected subnetwork for each of them. After adding about 10 reactions to a network of 50 reactions, the control coefficients become quite stable, so using a small subnetwork is justified here.

8.3 Resonances in metabolic control

Frequency-dependent control coefficients, as defined in section 7.2, describe the system's response to oscillatory perturbations. At certain frequencies, feedback loops in the network may lead to resonances where perturbations have large effects. If these oscillation periods are in the timescale of gene expression, then optimal expression should also be adapted to them. Mathematically, resonances reflect the eigenvalue spectrum of the Jacobian matrix $M^0 = N^0 \epsilon L$: its eigenvectors correspond to dynamical modes of the linearised system, and the eigenvalues determine the time-behaviour: vanishing eigenvalues indicate that the stationary state can be shifted, while negative real parts indicate exponentially decreasing modes, and non-vanishing imaginary parts indicate oscillatory behaviour. If the system is driven by oscillatory perturbations, the oscillatory modes, which usually

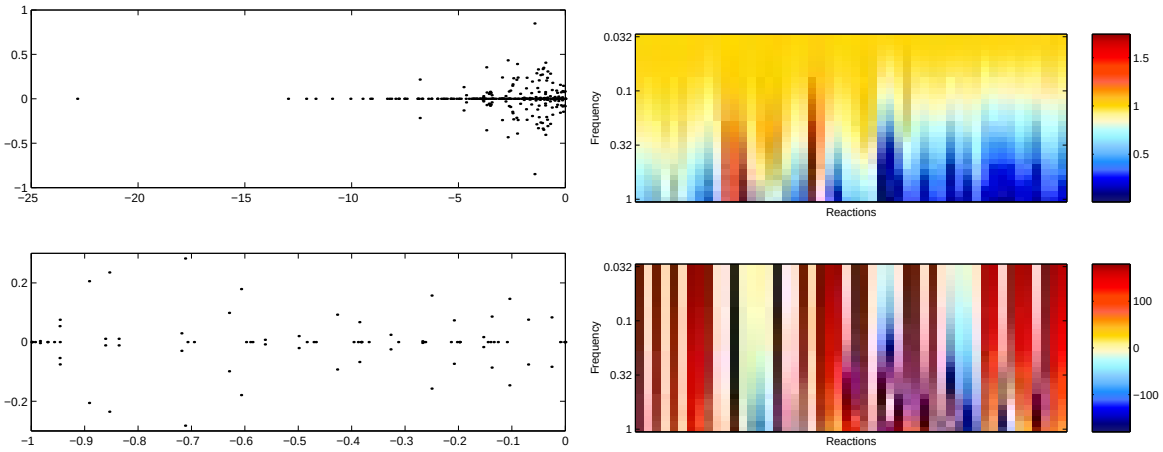


Figure 8.9: Frequency-dependent control coefficients. Left top: Eigenvalues of the Jacobian matrix M^0 shown as points in the complex plane. The real parts of the non-vanishing eigenvalues are negative. Left bottom: Only the eigenvalues near zero are shown. Eigenvalues with a real part close to zero can lead to a resonance at a frequency given by the imaginary part. Right: Control coefficients on Adenine at different excitation frequencies (on the ordinate). The abscissa refers to different reactions, sorted by their (stationary-state) control on Adenine. The absolute values (top) were normalised to the value at the lowest frequency, and phase angles (bottom) are expressed in degrees.

would decrease in time, lead to resonances which are especially strong if the real parts of the eigenvalue are close to zero, that is, just below a Hopf bifurcation.

Frequency-dependent control coefficients for the yeast metabolic network were calculated according to equation 7.15. As above, the elasticities of a reaction were simply assumed to be ± 1 for its substrates and products, and 0 otherwise. The system is stable, that is, there are no eigenvalues with positive real parts, and the oscillatory modes have only small real parts (see Figure 8.9, left), so they decay fast and no strong resonances occur. Figure 8.9, right, shows control coefficients on Adenine as a function of the frequency ω: the reactions are sorted by their stationary control, and only the reactions with the strongest control are shown. Control coefficients that are strong in the stationary case remain almost constant at higher frequencies, while the absolute values of smaller coefficients decrease further, and with some of them, the phase angles vary considerably. As there is no strong resonance in the control coefficients, also optimal gene expression need not depend strongly on the frequency here.

Chapter 9

Optimal expression and function

The model of optimal regulation claims a relation between optimal expression patterns and profiles of response coefficients. In this chapter, it is shown how qualitative and structural knowledge about the system to be regulated can be used to predict coregulation of genes, under the assumption of optimality. For instance, genes acting in functional modules are supposed to show low-dimensional, correlated expression patterns. For metabolic systems, sum rules for optimal expression patterns are derived from the theorems of metabolic control theory. Finally, expression patterns are compared directly to the simulated control coefficients.

9.1 Metabolic systems

9.1.1 Sum rules derived from metabolic theorems

Let us consider the optimal profiles (enzyme activities or the respective expression values) to achieve a required change $\mathrm{d}y$ of metabolic variables, for instance $\mathrm{d}y = \mathrm{d}\bar{y} - \mathrm{d}\hat{y}$ in the presence of a perturbation $\mathrm{d}\hat{y}$. If the output variables y describe metabolic fluxes or concentrations, then the theorems (1.16) of metabolic control theory lead to sum rules for the differential regulation profiles. In this section, the regulatory variables are supposed to describe enzyme concentrations E_i. The elasticity matrix π_E is considered invertible, which holds, for instance, if each enzyme catalyses exactly one reaction. According to equation 5.23, the optimal regulation profile fulfils

$$\mathrm{d}\bar{E}^* = (\pi_E^T)^{-1} F_{EE}\ \mathrm{d}\bar{E} = C^{y^T}\ (R_E^y\ F_{EE}^{-1}\ R_E^{y^T})^{-1}\ \mathrm{d}y \tag{9.1}$$

If the costs of different enzymes are independent of each other and if each enzyme catalyses exactly one reaction, then both F_{EE} and π_E^T are diagonal. In this case,

$\mathrm{d}\bar{E}^* = (\pi_E^T)^{-1} F_{EE}\ \mathrm{d}\bar{E}$ equals $\mathrm{d}\bar{E}$ up to a rescaling of the individual elements. The first term on the right-hand side of equation 9.1 is the transposed control coefficients matrix: so, like the metabolic flux distributions are linear combinations of flux modes (the columns of C^J), $\mathrm{d}\bar{E}^*$ is a linear combination of control profiles (the transposed rows of C^J).

If the output variables represent either only fluxes or only concentrations, then equation 9.1 leads to sum rules for $\mathrm{d}\bar{E}^*$:

1. If the fitness term $V(y)$ depends only on concentrations S, the summation theorem $C^S K = 0$ yields
$$\mathrm{d}\bar{E}^{*^T} K = 0 \tag{9.2}$$
For the proof, we transpose equation 9.1, postmultiply with K, and apply the summation theorem:
$$\mathrm{d}\bar{E}^{*^T} K = \mathrm{d}S^T (R_E^S\ F_{EE}^{-1}\ R_E^{S^T})^{-1}\ C^S K = 0 \tag{9.3}$$
Accordingly, $\mathrm{d}\bar{E}^{*^T} C^J$ vanishes as well.

2. If the fitness term $V(y)$ only depends on fluxes J, the connectivity theorem yields the sum rule
$$\mathrm{d}\bar{E}^{*^T} \epsilon L = 0 \tag{9.4}$$
because
$$\mathrm{d}\bar{E}^{*^T} \epsilon L = \mathrm{d}J^T (R_E^J\ F_{EE}^{-1}\ R_E^{J^T})^{-1}\ C^J\ \epsilon L = 0 \tag{9.5}$$
Similarly, we get $\mathrm{d}\bar{E}^{*^T} C^J = \mathrm{d}\bar{E}^{*^T}$ and $\mathrm{d}\bar{E}^{*^T} \epsilon C^S = 0$. These results resemble the statements for optimal enzyme concentrations derived in [67] under the constraint of a fixed sum of enzyme concentrations.

What is the meaning of the above sum rules? The first one, for the control of metabolites, implies that the elements of $\mathrm{d}\bar{E}^{*^T}$, summed over any stationary flux distribution, vanish. This holds, in particular, for the sum over any elementary mode [103]. As an example, let us consider the regulation of a metabolite in a unbranched chain where the stationary flux is described by $K = (1, 1, .., 1, 1)^T$. According to the sum rule 9.2, the scaled differential expression values in the chain sum to zero:
$$\sum_i (\pi_E)_{ii}^{-1} (F_{EE})_{ii}\ \mathrm{d}\bar{E}_i^T = 0 \tag{9.6}$$

The most efficient way to accumulate the metabolite is to upregulate the upstream enzymes and to downregulate the downstream enzymes.

The second rule, for the regulation of fluxes, predicts dependencies among the regulation patterns of neighbouring enzymes. If no conservation relations hold among the metabolites ($L = I$), then the i^{th} column of ϵL describes the reaction elasticities with respect to the i^{th} metabolite. If the reaction velocities depend only on concentrations of their own substrates and products, then all elements of the column vanish, except for the reactions of this metabolite. The sum rule 9.6 yields one linear equation for each metabolite: if the metabolite participates in n reactions (subscripted by i), then the scaled expression values $\mathrm{d}\bar{E}^*$ for the respective enzymes fulfil

$$\sum_i \mathrm{d}\bar{E}_i^{*T} \epsilon_i = 0 \tag{9.7}$$

In a series of experiments, the expression values of the n adjacent enzymes will be confined to an $(n-1)$-dimensional subspace. If a metabolite is involved in two reactions only, the ratio of the expression values $\mathrm{d}\bar{E}_i^*$ is fixed, that is, they are strictly correlated. In a unbranched reaction chain, each metabolite will usually have a negative and a positive elasticity on the producing and on the consuming reaction, so all enzyme changes will have the same sign and will be strictly correlated.

It is sometimes convenient to represent regulators, fluxes, and concentrations by logarithmic values. Then, the control coefficients have to be replaced by normalised control coefficients $\mathrm{dg}(J)^{-1}\, C^J\, \mathrm{dg}(J)$ and $\mathrm{dg}(S)^{-1}\, C^S\, \mathrm{dg}(J)$ in the above formulae. In addition, K and L have to be normalised by the stationary fluxes and concentrations, yielding $\mathrm{dg}(J)^{-1}K$ and $\mathrm{dg}(S)^{-1}L$.

9.1.2 Sum rule for the control of elementary flux modes

An additional sum rule can be derived for extreme flux distributions on the cone spanned by the elementary flux modes (see section 1.4.2. Let us consider a nonlinear fitness function with a local optimum outside the cone, as shown in Figure 9.1, right. The optimum, constrained to the cone, lies on a boundary spanned by some of the elementary modes k_i. What happens to this optimum if the fitness function is changed? For small perturbations, it is improbable that the fitness optimum crosses the cone boundary or that the constrained optimum jumps to a different boundary. Usually, the optimal flux J will remain inside the same boundary. Let the matrix K_B contain, in its columns, linearly independent modes spanning the respective boundary, and let the columns of a matrix $K_\perp$ span the subspace orthogonal to the boundary. The space of (both stationary and non-stationary) flux distributions is then spanned by $(K_B|K_\perp)$. As the flux change $\mathrm{d}\bar{J}_\perp$ orthogonal to the boundary must vanish, the regulation profile fulfils the sum rule

$$\mathrm{d}\bar{J}_\perp = K_\perp^T R_x^J \mathrm{d}\bar{x} = 0 \tag{9.8}$$

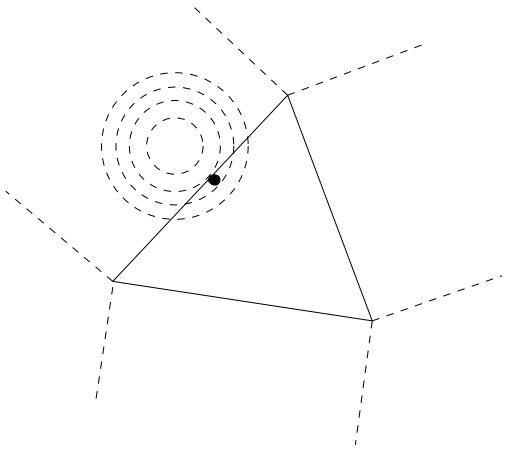


Figure 9.1: Regulation of elementary flux modes. Right: The cone spanned by three irreversible elementary flux modes (compare Figure 1.4) is shown as a triangular section. A fitness function (contour lines shown by dashed circles) has its optimum outside the cone. The optimum constrained to the cone lies on a face of the cone (dot). A global optimum within the dashed triangle next to a corner would lead to a constrained optimum on the corner. If the fitness landscape is changed and the optimum moves by a small distance, the constrained optimum rests inside the same face, and the optimal flux change $\mathrm{d}\bar{J}_\perp$ orthogonal to the boundary will vanish.

9.1.3 Optimal response to flux perturbations

If $\pi_E = I$, that is, if control and response coefficients are identical, then the optimal response to a stationary flux perturbation $\mathrm{d}\hat{J} \in M_J = \mathrm{span}(K)$ (compare Figure 1.4) reads

$$\begin{aligned} \mathrm{d}\bar{E} &= -(F_{EE} + R_E^{J^T} F_{JJ} R_E^J)^{-1} R_E^{J^T} F_{JJ} \mathrm{d}\hat{J} && (9.9) \\ &= -(F_{EE} + C^{J^T} F_{JJ} C^J)^{-1} C^{J^T} F_{JJ} C^J \mathrm{d}\hat{J} && (9.10) \end{aligned}$$

as $\mathrm{d}\hat{J} = C^J\, \mathrm{d}\hat{J}$. If F_{EE} is isotropic, this can be interpreted geometrically: as $C_E^{J^T} F_{JJ} C_E^J$ is a symmetric operator, the unregulated flux change $\mathrm{d}\hat{J}$ can be expanded into a sum $\mathrm{d}\hat{J} = \sum_i \mathrm{d}j_i$ of orthogonal eigenvectors $\mathrm{d}j_i$ of $C_E^{J^T} F_{JJ} C_E^J$ with eigenvalues λ_i. Each $\mathrm{d}j_i$ lies either in $M_S = \mathrm{span}(\epsilon L)$ (if $\lambda_i = 0$) or in the orthogonal space $M_S^\perp$. Equation 9.10 yields

$$\mathrm{d}\bar{E} = \sum_i -\frac{\lambda_i}{F_{EE} + \lambda_i} \mathrm{d}j_i \qquad (9.11)$$

As F_{EE} and the λ_i are both negative, the coefficients for the different eigenvectors vary between -1 (for $|F_{EE}| \ll |\lambda_i|$) and 0 (for $|F_{EE}| \gg |\lambda_i|$). For $F_{EE} \to 0$, they can only assume the values -1 and 0, so the mapping from $\mathrm{d}\hat{J}$ to $-\mathrm{d}\bar{E}$ becomes an orthogonal projector to $M_S^\perp$.

9.2 Functional modules

The statistical features of the response coefficients reflect, among other things, the system's large-scale structure. If the cell contains specialised subsystems [63], such as protein complexes or metabolic subnetworks maintaining particular metabolic fluxes, then the response coefficients matrix be decomposed into a product $R^y_x = R^y_c \; R^c_x$ where R^c_x has a block structure. R^c_x describes the influence of individual regulators on the parameters or variables that are directly influenced by the module, while their influence on the output variables y is described by R^y_c. If F_{xx} is isotropic, equation 5.30 implies that regulators from the same modules are coregulated: for instance, if the proteins form complexes and if each protein belongs to one complex only, they will have proportional differential expression patterns, that is, their linear correlation is ± 1. Equation 5.30 has, again in the isotropic case, also a consequence for modules or complexes that affect only some of the output variables: if a module of n regulators affects only $m < n$ of the output variables, its differential expression patterns will be confined to an m-dimensional subspace.

9.2.1 Cooperation between modules

If several modules of regulators influence the same cell variables, they can cooperate or share their work, so we may expect that their behaviour is effectively coupled by the optimality postulate. It will be shown in the following that the optimal regulation by the cooperating modules can be expressed by the local behaviour of the modules observed in isolation, in analogy to modular response theory [8]. How is the optimal behaviour of a module affected by the presence of other modules? Let us consider a single module $x^{(i)}$ which, given an external perturbation $\mathrm{d}\hat{y}$, would contribute a change

$$\mathrm{d}y^{(i)} = R^y_x \mathrm{d}\bar{x}^{(i)} = A^{(i)} \; \mathrm{d}\hat{y} \tag{9.12}$$

Now let us suppose that several modules $x^{(k)}$ are present, and each of them behaves optimally given the optimal behaviour of the other modules. What is then the optimal contribution $\mathrm{d}y^{(i)}$ of module i? The optimal $\mathrm{d}\bar{y}$ for a given perturbation $\mathrm{d}\hat{y}$ can be decomposed into

$$\mathrm{d}\bar{y} = \mathrm{d}\hat{y} + \sum_k R^{y^{(k)}}_x \mathrm{d}\bar{x}^{(k)} = \mathrm{d}\hat{y} + \sum_k \mathrm{d}y^{(k)} \tag{9.13}$$

If the other modules already behave optimally, the remaining perturbation seen by module i reads

$$\mathrm{d}\hat{y} + \sum_{k \neq i} \mathrm{d}y^{(k)} = \mathrm{d}\bar{y} - \mathrm{d}y^{(i)} \tag{9.14}$$

The optimal response of module i is known: it is given by equation 9.12. Inserting the perturbation seen by module i as $\mathrm{d}\hat{y}$ into equation 9.12 yields

$$\begin{aligned} \mathrm{d}y^{(i)} &= A^{(i)}(\mathrm{d}\bar{y} - \mathrm{d}y^{(i)}) \\ \Rightarrow \mathrm{d}y^{(i)} &= (I + A^{(i)})^{-1}A^{(i)}\ \mathrm{d}\bar{y} = B^{(i)}\mathrm{d}\bar{y} \end{aligned} \tag{9.15}$$

Equation 9.13 thus yields

$$\begin{aligned} \mathrm{d}\bar{y} &= \mathrm{d}\hat{y} + \sum_k B^{(k)}\mathrm{d}\bar{y} \\ \Rightarrow \mathrm{d}\bar{y} &= (I - \sum_k B^{(k)})^{-1}\mathrm{d}\hat{y} \\ \Rightarrow \mathrm{d}y^{(i)} &= B^{(i)}(I - \sum_k B^{(k)})^{-1}\mathrm{d}\hat{y} \end{aligned} \tag{9.16}$$

The last equation shows that the behaviour of modules in the complete system can be expressed by the local behaviour of the modules in isolation, contained in $B^{(i)}$, which becomes coupled by the matrix inverse. The optimal regulation profile $\mathrm{d}\bar{x}^{(i)}$ can be obtained from $\mathrm{d}y^{(i)}$ by equation 5.23.

9.3 Relating expression to control coefficients

Equation 9.1, rewritten for regulators x, states that optimal expression and control coefficients are related by

$$\mathrm{d}x^* = (\pi_x^T)^{-1}F_{xx}\ \mathrm{d}x = C^T(R_x^y F_{yy} R_x^{y^T})^{-1}\mathrm{d}y \tag{9.17}$$

where the matrix C contains the control coefficients on the fluxes and concentrations that are relevant for the fitness. If $F_{xx}^{-1}\pi_x^T$ is diagonal, then the expression matrix from a series of experiments is supposed to have the form

$$X = D\ C^T\ M \tag{9.18}$$

where the diagonal matrix D scales the values for the individual genes. Except for this scaling, expression profiles should consist of superposed profiles of control coefficients. An equation of the same form as 9.18 holds for normalised control coefficients, describing the relation between logarithmic expression values and fluxes or concentrations. The ICA modes estimated from experimental data support this hypothesis by their qualitative properties: like the control coefficient profiles, independent expression components were localised in the metabolic network, and some of them even represented metabolic subsystems.

In this section, experimental expression data will be related directly to simulated control coefficients. First, it is tested whether the data support relation 9.18 at all, that is, whether expression data and control coefficients show significant similarities. Secondly, common independent components behind the both kinds of data are determined. Besides the control coefficients themselves, also their absolute values, and log-transformed absolute values are considered.

The flux control coefficients C^J were calculated as described in section 8.2, and ORF and chemical reactions were made comparable by mapping them to EC numbers (see Appendix B.4). For the comparison, the expression data were projected to their first principal components to reduce the noise. Similarly, instead of specifying which of the variables y are relevant for the fitness, I chose the first few principal components of the control coefficients which are likely to capture the control on many "candidate" fluxes. Except for the centring, these principal components equal the uncorrelated control profiles from chapter 9. Thus the logarithmic expression data X and the (possibly transformed) flux control coefficients C^{J^T} are represented by their first principal components according to

$$X \approx \tilde{X}^{(n_x)} A_x^{(n_x)} \tag{9.19}$$

$$C^{J^T} \approx \tilde{C}^{(n_c)} A_c^{(n_c)} \tag{9.20}$$

where the matrices $\tilde{X}^{(n_x)}$ and $\tilde{C}^{(n_c)}$ contain the first n_x and first n_c principal components of the expression data and the control coefficients matrix, respectively. Thus for the tests, equation 9.18 is replaced by

$$\tilde{X}^{(n_x)} \approx D\ \tilde{C}^{(n_c)}\ M \tag{9.21}$$

The relation 9.21 is tested quantitatively in two ways: (1) D and M are fitted by maximum likelihood. (2) Similarities between $\tilde{X}^{(n_x)}$ and $\tilde{C}^{(n_c)}$ are quantified by angles between the subspaces spanned by their columns. The significance of the results is studied by permutation tests[1].

What is the meaning of the matrix D? To calculate optimal expression patterns, the response coefficients $R_x^y = C^y \pi_x$ have to be known, rather than the control coefficients C^y themselves. If each enzyme (or gene) acts on a different reaction, then π_x is diagonal, and response and control coefficients differ only by a reaction-specific factor. The matrix D describes, among other things, the signs of these elasticities, which cannot be inferred from the metabolic network unless the directions of the metabolic fluxes are known.

[1]Let us consider a suspected relation between the columns of A and B, quantified by a test statistics $s(A, B)$. The relation is not supposed to hold between A and any arbitrary matrix, such as randomised versions B_{rand} of B, in which the elements have been permuted randomly within the columns. We shall conclude that the relation holds for A and B if $s(A, B)$ differs significantly from realisations of $s(A, B_{rand})$, and the test statistics s is characterised by a p-value against its values for randomised matrices B.

9.3.1 Explaining expression data by control coefficients

Given the matrices $\tilde{X}^{(n_x)}$ from experimental expression data and and $\tilde{C}^{(n_c)}$ from simulated control coefficients C^J, the matrices D and M in equation 9.21 were fitted[2]. Figure 9.2, left, shows the results from three expression data sets (Gasch et al. [32], Causton et al. [10], Spellman et al. [107]), for $n_x = 2, 4, 6, 8, 10$ and $n_c = 10$ principal components. The quality of the fit, measured by the fraction of variance of $\tilde{X}^{(n_x)}$ explained by the right hand side of equation 9.21, is shown by red dots, while the blue error bars show the same for randomised data X. Only for the cell cycle data, the results are significant at 5% level (indicated by stars). In a second approach, $D = I$ was kept fixed, and the simple regression model $\tilde{X}^{(n_x)} = \tilde{C}^{(10)}\ M$ was fitted (9.2, centre). With absolute and log-absolute control coefficients, the fit is significantly good at 5% level for all data sets and n_x studied. If in addition, an appropriate sign is chosen for each gene (according to the elements of D estimated before), the variance explained increases again. Significance was not studied for this case.

9.3.2 Similarity between gene expression and control coefficients

Alternatively, the similarity between $\tilde{X}^{(n_x)}$ and $\tilde{C}^{(n_c)}$ was measured by the angle between the subspaces spanned by their columns[3]. The angles were calculated for different subspace dimensionalities n_x and n_c and scored by the p-value from a permutation test as described above. The matrix of p-values $p(n_x, n_c)$, for different choices of (n_x, n_c), is shown in the top row of Figure 9.3 (cell cycle data [107]). Results for the environmental changes data Causton et al. [10] are shown in Appendix C.

For absolute and logarithmic control coefficients, many of the p-values are small: the cumulative histograms $\mathcal{F}(p)$ (shown by black lines in the second row of the figures) clearly differ from the straight line expected for a uniform distribution. However, it is not obvious if the whole result is significant, because the p-values for different n_x and n_c represent multiple tests, and they are, in addition, dependent. To study the significance of the matrix of p-values as a whole, a permutation test was applied to the cumulative distribution $\mathcal{F}(p)$ of the p-values in the matrix. In the second row of Figure 9.3, the cumulative distributions $\mathcal{F}(p)$ of the p-values (black) are compared to the cumulative distributions obtained from a permutation test (red lines), where the whole analysis had been repeated

[2]Practically, they were estimated iteratively by maximum likelihood.

[3]For centred statistical variables, the cosine of the angle equals the linear correlation between them. For spaces spanned by several statistical variables, the angle describes the correlation between the first canonical variables.

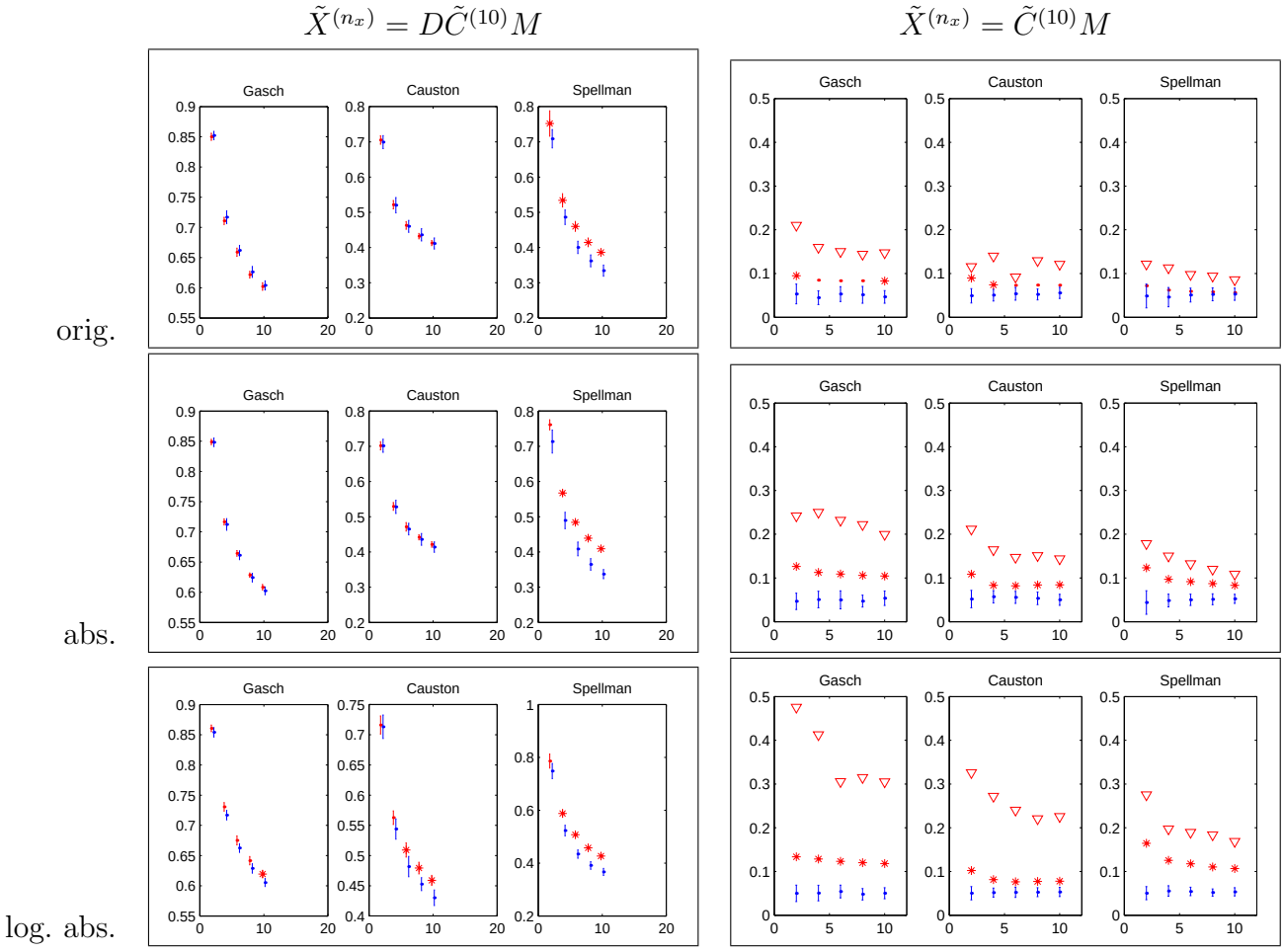


Figure 9.2: Expression data are related to control coefficients. According to equation 9.21, the first $n_x = 2, 4, 6, 8, 10$ principal components $\tilde{X}^{(n_x)}$ of expression data were explained by the first ten principal components $\tilde{C}^{(10)}$ of simulated flux control coefficients. Left boxes: The diagonal matrix D and a matrix M were estimated. The large boxes correspond to different transformations for the control coefficients: (1) original values (2) absolute values (3) logarithms of absolute values. Each small box shows the fraction of data variance explained (stars) for one expression data set (from [32] [10] [107]) and different numbers n_x of components. The results for randomised matrices are shown as dots with blue error bars (mean and standard deviation). The red error bars represent repeated estimation runs, and stars indicate significant results (at 5% level). The cell cycle data from Spellman et al. are significantly well explained by the flux control coefficients. Right: Same, with fixed $D = I$. For absolute and log. absolute control coefficients, all results are significant at 5% level. Triangles: fixed diagonal elements ± 1 were chosen for D. The signs were obtained from the estimations shown on the left, and M was estimated. The fit becomes much better, but significance was not assessed here.

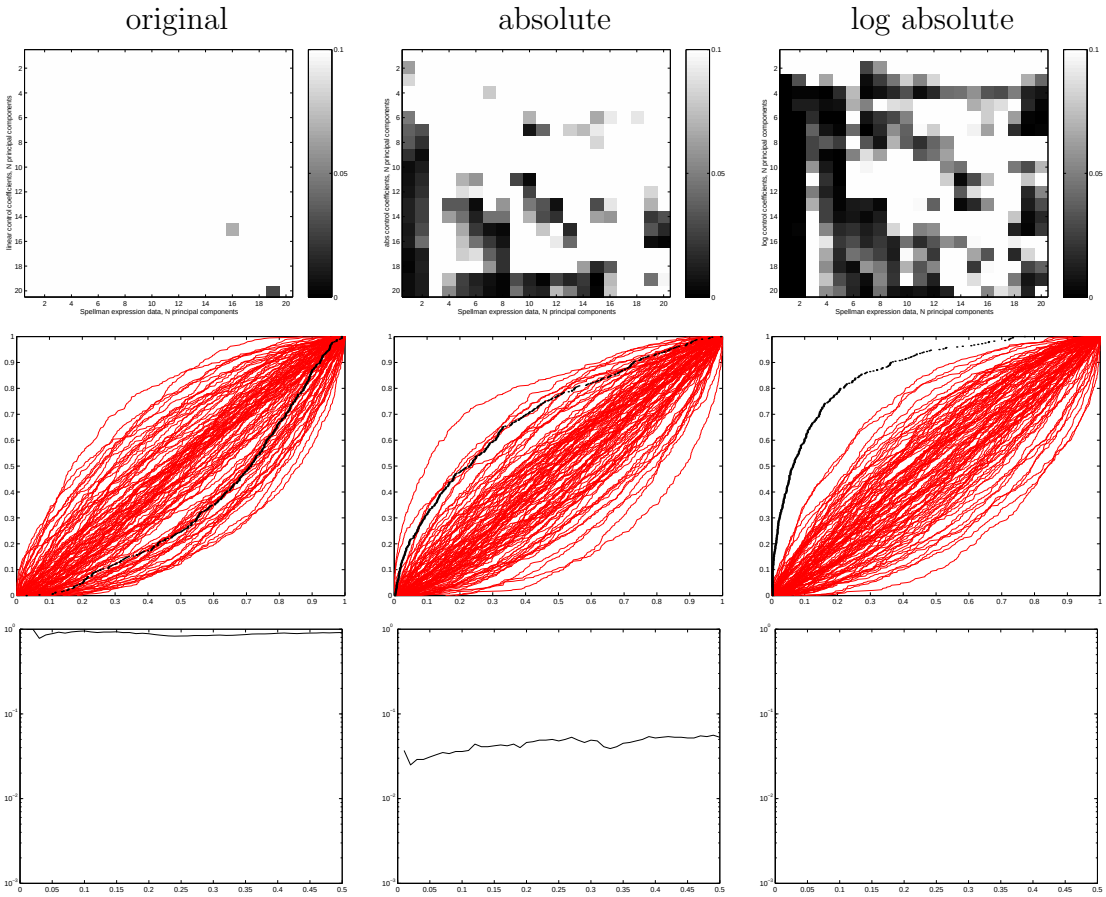


Figure 9.3: Similarity between gene expression data and control coefficients. Expression data X (cell cycle data [107]) and simulated control coefficients C^J are compared by calculating the angle between the subspaces spanned by their first n_x and n_c principal components. These angles were calculated for different subspace dimensionalities (n_x, n_c) and scored by p-values, obtained by a permutation test. Top: Matrix of the p-values (colour-coded) for different pairs of n_c (matrix rows) and n_x (matrix columns). The results for original flux control coefficients (left), as well as their absolute values (centre) and logarithmic absolute values (right) are shown. For the last two cases, many small p-values are present. Centre row: To test this result for significance, the cumulative density $\mathcal{F}(p)$ of p-values (black curves) is compared to the distributions that would be expected by chance (red curves, from a permutation test where the whole analysis was repeated for randomised expression data). Bottom: the cumulative density $\mathcal{F}(p)$ is characterised again by p-values $P(p)$. Abscissa and ordinate refer to p and $P(p)$, respectively. With the log-absolute control coefficients (right), all p-values $P(p)$ are below 0.001.

for randomised matrices X. For each p-value p, the observed value $\mathcal{F}(p)$ can again be characterised by a p-value $P(p)$ (shown in the third row). For log-transformed control coefficients (Figure 9.3, right), there are significantly many small p-values with the cell cycle data Spellman et al. [107] (Figure 9.3), the cell stress data Causton et al. [10] (Figure C.5), and the cell stress data Gasch et al. [32] (not shown). The results indicate a significant similarity between expression data and the log-transformed absolute control coefficients. For the original control coefficients, no significant similarity has been found. Thus the expression data can be explained by a common activation of genes with a strong control on the dominant uncorrelated flux distributions, but not by the regulation of the same genes according to their control profiles, where the signs of the elasticities π_x are neglected.

9.3.3 Common components behind gene expression and control

If the expression profiles are linear combinations of control coefficient profiles, both kinds of data should contain common components. To identify such components, ICA was applied to the matrix $(\tilde{X}^{(6)}|\tilde{C}^{(12)})$ combined from principal components of the expression data and the control coefficients[4], yielding a decomposition

$$X \approx Q^T \; B_x \; A_x^{(6)} \tag{9.22}$$

$$C^{J^T} \approx Q^T \; B_c \; A_c^{(12)} \tag{9.23}$$

Setting $K = (B_c \; A_c^{(12)})^T$, C^J is decomposed into $C^J \approx KQ$ as in chapter 8, but with a different criterion: here the rows q_i of Q are common components for both expression and control profiles, and they are supposed to be statistically independent. Each of these control profiles represents genes that respond to an expression mode and control a metabolic flux mode. The time series of the expression mode is given by the corresponding row of the loadings matrix $Y = B_x \; A_x^{(6)}$, and the flux mode is contained in the respective column of K.

Combined metabolic/expression components were determined from the stress response data Gasch et al. [32], which had already been studied in chapter 3 and section 7.1.

[4]Log-transformed expression data from the environmental changes experiments Gasch et al. [32] were averaged to yield effective expression values for the EC numbers. The resulting 186×174 matrix was centred and decomposed by PCA into a product $X = \tilde{X} A_x$. The first 6 principal components are contained in a matrix $\tilde{X}^{(6)}$. The absolute values of simulated flux control coefficients were log-transformed. The rows of the transposed matrix were averaged over reactions with the same EC numbers. The resulting 186×566 matrix was also centred and decomposed by PCA into a product $C^{J^T} = \tilde{C} A_c$. The first 12 principal components are contained in a matrix $\tilde{C}^{T(12)}$. The combined matrix $(\tilde{X}^{(6)}|\tilde{C}^{(12)})$ was decomposed by ICA into a product $Q^T(B_x|B_c)$. Twelve independent components q_i (columns of Q^T) were estimated ten times with different random seeds, averaged, and sorted with respect to the variance of the expression data which they explain.

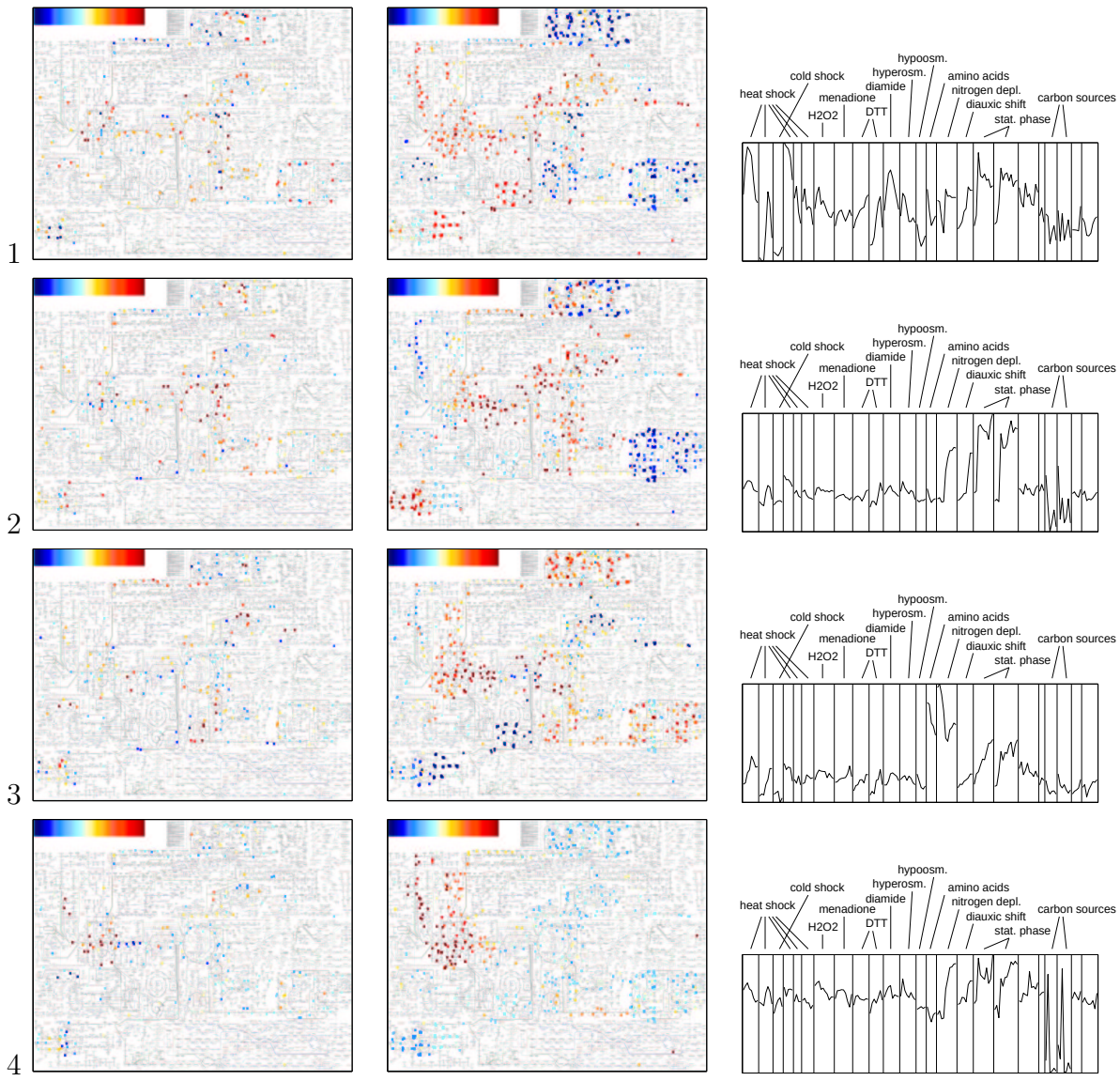


Figure 9.4: Independent components behind both expression data [32] and flux control coefficients, according to equations 9.22 and 9.23. Top left: first independent component q_1. Top centre: Loadings for the control coefficients (first column of K). Top right: Loadings for the expression data (first row of Y). Component 1 resembles the so-called "environmental stress response", the largest gene cluster determined in [32]. The other rows of diagrams show the components 2-4. Component 2 is partly related to the citric acid cycle. Component 3 probably describes the adaptation to the minimal medium used in the amino acid starvation and nitrogen depletion experiments. Component 4, which activates galactose degradation and upper glycolysis, is activated in yeast grown on galactose. The components 5-12 are shown in the Figures 9.5 and 9.6

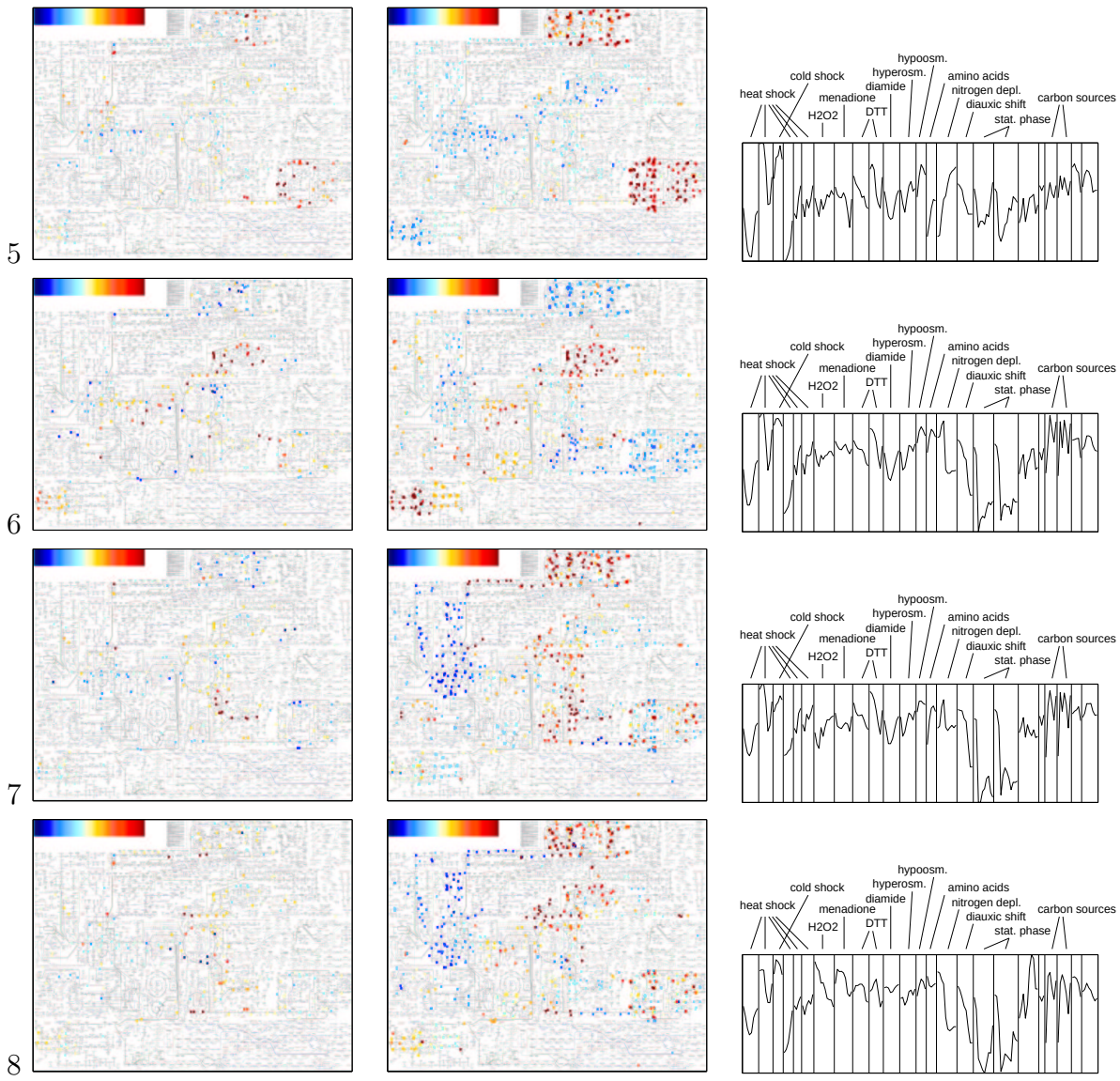


Figure 9.5: Independent components behind both expression data and flux control coefficients (see Figures 9.4 and 9.6). Here the components 5-8 are shown. Component 5 corresponds to the synthesis of nucleotides while component 6 is involved in amino acid synthesis. Component 7 is related to the pathway leading to proline.

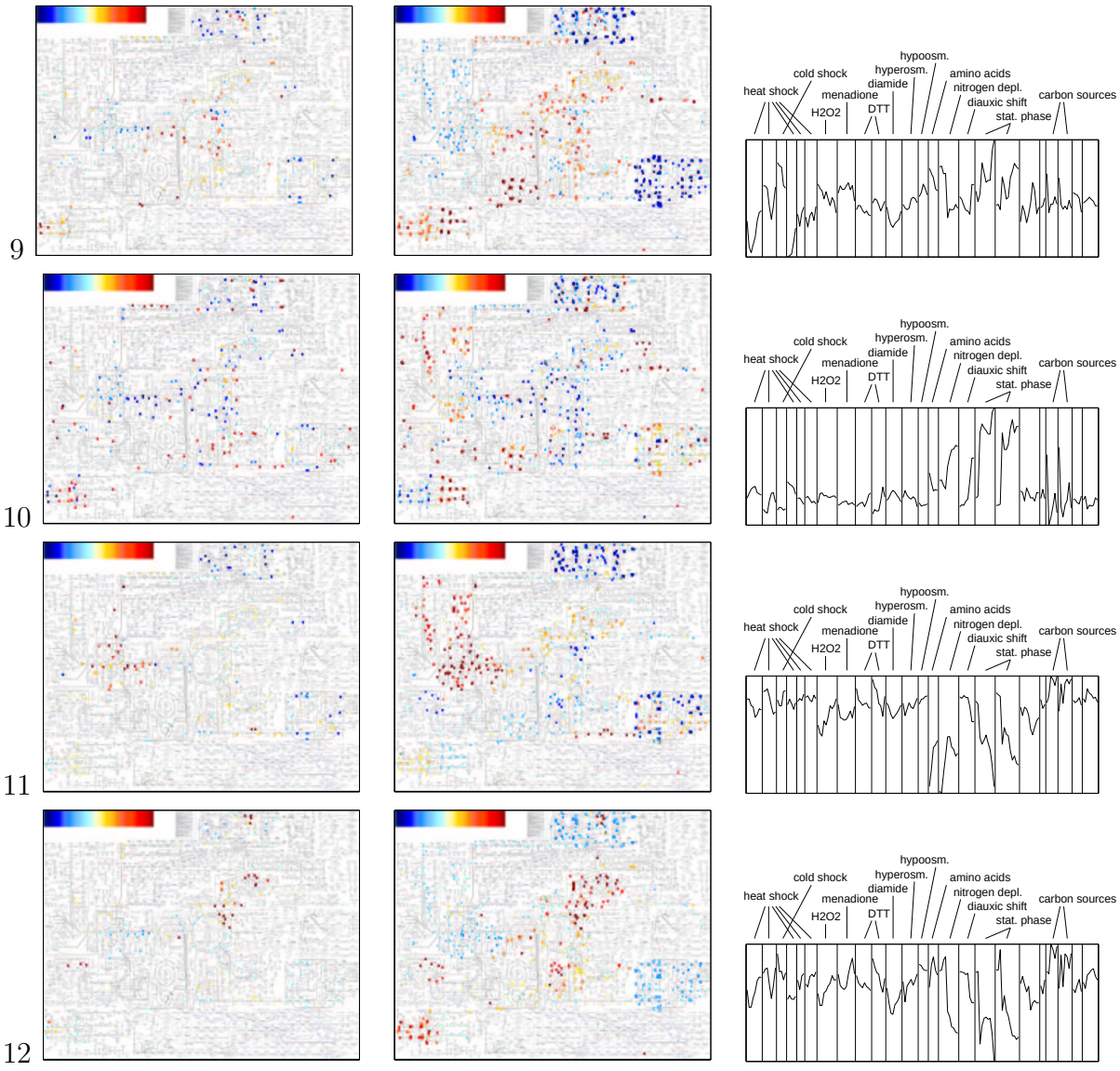


Figure 9.6: Independent components behind both expression data and flux control coefficients (see Figures 9.4 and 9.5). Here the components 9-12 are shown. Component 9, which activates the citric acid cycle and inhibits glycolysis, is upregulated in growth on ethanol and at the end of the diauxic shift. Component 11 activates glycolysis and inhibits nucleotide synthesis and is downregulated during the diauxic shift.

The data describe the adaptation of yeast to different stress conditions and to restricted nutrient supply. The 12 estimated components, which explain about 62% of the variance in the expression data, are shown in the Figures 9.4, 9.5, and 9.6.

Each component q_i and the corresponding absolute flux mode k_i are plotted on the Boehringer chart, while the corresponding expression mode y_i (row of $Y = B_x\ A_x^{(6)}$) is shown as a time series. Most of the control profiles correspond to specific parts of the metabolism, such as the citric acid cycle, the degradation of glucose and galactose, and the production of nucleotides and different amino acids. The expression modes show how these parts are regulated in the different experiments: for instance, the citric acid cycle, representing the second component, is strongly activated by nitrogen depletion, during the diauxic shift and growth on Ethanol, and in the stationary phase. In addition, it shows minor responses within some of the stress reactions.

The adaptation to different stress conditions is supported by global metabolic changes, which were visualised here by integrating expression data and control coefficients profiles. In the original work [32], the experimental data had been analysed by hierarchical clustering, and accordingly, the results were described by a stereotypic response to cell stress shown by many genes and by specific responses to certain stress conditions shown by smaller gene clusters. Here, on the contrary, a linear superposition of expression modes allowed for disentangling different metabolic processes present in all reactions, but to a different extent. This method of integrating expression data and control coefficients is in line with the predictions made from optimal expression, and furthermore, it is justified by the significant relations between expression and control coefficients found in the previous sections.

9.4 Discussion

Gene expression and metabolic control coefficients. According to the model of optimal differential expression, genes with similar response coefficients should be coregulated. Empirically, coregulation had been reported previously for genes within functional classes [111], interacting proteins [36], and proteins forming permanent complexes [62]. These experimental results are consistent with the present predictions if such functional gene groups are characterised by similar response coefficients, and this is probably the case: gene annotations that are based on studies of mutants state that a particular cell function responds strongly to the loss of a gene, implying a high value of the response coefficient. On the other hand, proteins in permanent complexes or proteins forming permanent interaction pairs can be assumed to have similar response coefficients: if only the complex as a whole is functional, then underexpression of any of its constituents will have

similar results for the cell, thus the response coefficients with respect to the constituents should at least be proportional.

The aim of this chapter was to predict and to test quantitative relations between gene expression and control coefficients. In contrast to other studies, where expression was related to functional gene categories [111] [7] or heuristical definitions of metabolic pathways such as KEGG metabolic pathways [71] or shortest paths [118] [38] [117], the present approach describes the function of genes by an established quantitative concept, namely the response coefficients defined in metabolic control theory. Thus I could refer to the extensive theoretical results from metabolic control theory, in particular, to the summation and connectivity theorems for metabolic control coefficients: they were used here to derive sum rules that relate optimal expression profiles to the structure of the metabolic network.

Experimental expression data from the cell cycle and cell stress experiments were compared to metabolic control coefficients. The biological idea was that both cell cycle processes and stress reactions require particular material resources, and thus should be accompanied by global adaptations of metabolism, concerning large parts of the metabolic network. The fitness function was supposed to depend on metabolic fluxes, but was not specified in detail: the output variables to be regulated were, effectively, the first uncorrelated flux modes, as defined in chapter 9. An accurate comparison would require the response coefficients of chemical reaction with respect to the gene expression values: as these were not available, the response coefficients were calculated under very simplified assumptions. These simulated coefficients can only serve as preliminary tests because their exact values are not trustworthy, but still, significant relations to the expression data have been found.

The control coefficients for a large metabolic system were calculated from the topology of the metabolic network alone, with simple assumptions about the reaction elasticities. Furthermore, ORF and reactions were only matched via the (non-unique) EC numbers. Calculating control coefficients from the yeast network published in [30] may overcome this problem. Thirdly, it further assumed that the response coefficients could be replaced by control coefficients, that is, the elasticities π_x between expression values and reaction velocities were neglected. In reality, the signs of control and response coefficients differ: if the metabolic flux of a reaction is negative, all the response coefficients of the enzyme will carry an additional negative sign which I did not account for here. The elasticities π_x together with their signs were estimated (see Figure 9.2) from the comparison of expression data and control coefficients, which worked well for the cell cycle data. For the other data sets, the relation between expression and control coefficients became insignificant, probably due to the high number of parameters estimated. On the contrary, significant relations between expression data and the absolute control coefficients (with and without

log-transformation) were found for all data sets.

Knowing that there is a significant relation, common components between the stress response data Gasch et al. [32] and the log-absolute control coefficients were determined by ICA. The integration of expression data and information about the metabolic network yields combined metabolic/expression components, which gave a vivid picture of the metabolic processes that assist the adaptation to environmental changes.

Summarising the predictions for gene expression profiles. At present, the analysis of optimal differential expression cannot be used for predicting real gene expression patterns, because only few response coefficients are known for appropriate systems and also the fitness function can only be assumed. But still, partial knowledge about the system to be regulated can be used to predict general properties of expression patterns. The predictions made in this work are summarised in the following: first, the linear structure of the model implies linear dose response curves and linear superposition of the responses to different perturbations, but linearity alone could also follow from a causal model of regulation. Additivity of expression patterns during heat shock and hypo-osmotic shock was experimentally shown in [32] (web supplement). The optimality assumption requires that the response is an appropriate answer to the perturbation and that it damps fitness losses. A perturbation is distributed by a cascade of responses and may affect subsystems which do not seem directly concerned: for instance, if the response to a perturbation requires energy, then energy production should be coactivated.

Expression patterns reflect the response coefficients on the relevant variables: if the fitness curvatures F_{xx} are isotropic, then the differential expression pattern after a perturbation of cell variables is a linear combination of response coefficient profiles. Even if the response coefficients are not known, this allows for qualitative predictions: genes that have no effect on the concerned variables remain unchanged. Genes with similar functions, in particular cooperating genes, are coregulated, and superfluous gene products are downregulated so that resources can be allocated to other processes. These assertions are backed by expression data from several experiments (see, e.g., [18] [32] [10]). According to the findings in chapter 8, the genes' response coefficients on metabolic fluxes and concentrations are almost sparse and concentrated in the metabolic network, thus linear components with the same properties may appear in expression data, such as the independent components found in section 4.3.4. Relations between expression profiles and the structure and kinetics of the metabolic network were stated by sum rules in section 9.1. Cooperating gene products should show correlated expression, and groups of regulators with a direct influence on few cell variables should show low-dimensional expression patterns. In deletion experiments, genes should compensate for each other in a symmetric way, as shown in section 5.3.3. Genes that are activated after a change of environmental conditions should also become more important for growth.

Correspondence between expression and function. The present analysis of optimal regulation claims that regulation profiles, and also the structure of the regulatory machinery, tend to portray aspects of the system to be regulated. This correspondence can be attributed to evolution: if a gene is supposed to maximise the organism's fitness under typical evolutionary conditions, its marginal fitness $G_x = \nabla_x G = F_x + F_y R^y_x$ must vanish despite any perturbation. The marginal fitness reflects the response coefficients, so the evolving gene program implicitly stores information about the functional structure of the cell. In experiments, this information can be read by probing the regulatory system with perturbations.

Regulatory systems which reflect their objective are well-known in biology: a striking example is with operons, where sets of cooperating genes are controlled by the same transcriptional machinery. Another example can be found in the regulation of amino acid synthesis: the aspartate kinase is the first enzyme in the pathway for the synthesis of threonine, isoleucine, lysine, and methionine. The three isoenzymes AspKI, AspKII, and AspKIII receive negative feedback signals from the amino acids, thus portraying the strong control of aspartate kinase on amino acid levels. This pattern of regulation even appears on two levels, as the feedback signals are realised by both allosteric inhibition and repression of gene expression (see [77])

Isoenzyme	Allosteric inhibitor	Genetic repression by
AspK I	Thr	Thr, Ile
AspK II	-	Met
AspK III	Lys	Lys

A similar correspondence appears also in image-processing: valuable information in pictures can be enhanced (e.g., for the segmentation of textured images) by a convolution with linear filters. For collections of images, optimal linear filters can determined, which will in turn reflect typical signals, such as edges or objects of a particular scale of length. I shall finish this thesis by citing J. W. Goethe, who expressed this correspondence between recognition and objects to be recognised in a poetic and most condensed manner:

Wär nicht das Auge sonnenhaft, die Sonne könnt es nie erblicken[5].

[5]If the eye were not sun-like, it could not see the sun.

Chapter 10

Conclusions

The studies in this thesis show that (1) linear components extracted from several sets of gene expression data reflect biological functions of the genes and that (2) this finding is consistent with the hypothesis that expression patterns are shaped by optimality.

Linear models. Gene expression data contain coregulation structures that are not immediately visible in the data matrix and thus require statistical analysis. While clusterings determine groups of genes with similar expression profiles, factor models explain each gene profile by an individual superposition of basic profiles. In this thesis, expression data were analysed using independent component analysis and other factor models. The basic profiles, called "expression modes", were determined blindly, according to different statistical criteria. Coregulation structures were visualised by plotting the different expression modes and the corresponding components, and thresholding of particular components was used to identify genes that respond strongly to the modes. A model for the simulation of gene expression data was proposed: in this model, the genes are regulated by hidden variables ("biological expression modes"), and each gene responds to them individually according to a nonlinear function, the so-called gene program. Depending on the interpretation, the modes may represent biological signals or global variables parametrising the cell state. The model accounts for the superposition of effects in expression data, and after linearisation, it has the same form as the statistical models used for data analysis.

Data analysis. Is there a chance to resolve biological modes with blind data analysis? The tests with simulated expression data show that the modes can be found with ICA quite reliably, also if the data are noisy and the numbers of components do not exactly match. Weak nonlinearities in the data could be reduced by a proper preprocessing and play a minor role in the reconstruction of modes. Different statistical methods were tested on simulated and experimental expression data: the tests included supervised and unsupervised estimation of the gene programs, and the results were confirmed by cross-validation. To test whether the expression of genes could be explained by the action

of transcription factors, regression models were fitted to data for individual genes: the gene profiles were alternatively explained by global modes estimated by factor models and by the expression of specifically chosen transcription factors. The global modes yielded better predictions.

Nonlinear ICA was applied to yeast expression data from the environmental changes experiment [32]: the expression profiles could be explained by few components, and the estimated gene programs show moderate nonlinearities. Whole-genome data from cell cycle experiments [107] were analysed in more detail with linear methods that are sensitive to sparse components. Some of the modes show oscillatory time series and could be related to cell cycle processes, and also presumable experimental artefacts could be detected by ICA modes. The components from different methods span similar data subspaces, but the individual components differ among the methods: there are, however, noteworthy similarities, for instance between ICA and k-means clustering. With the cell cycle data, the results of FastICA were robust against a resampling of genes and samples. This indicates that the modes found are active in more than a few samples and affect many genes: contrariwise, it might be difficult to find modes with specific influences on a few genes. Other authors have shown that coregulated genes, determined by clusterings, tend to be functionally related. Their findings were confirmed in this work, but based on a different concept of coregulation: genes were considered coregulated if they responded strongly to particular expression modes. For several expression modes, the corresponding "differentially expressed" genes share biological functions, which was shown by plotting them on the metabolic map or by comparing them to functional categories from MIPS.

Optimality. If gene products cooperate, it is plausible that they should be coregulated. As the functional efficiency of the cell requires appropriate expression patterns, evolution will shape the regulatory machinery: genes should sense signals that provide the most valuable information about the necessity of expression. From a teleological point of view, the target processes of a gene could even be seen as a cause, the causa finalis, of its expression. To formalise this approach, a principle of optimal regulation was proposed in this thesis. The corresponding model of regulation describes the optimal response of regulators to small perturbations a stationary cell state. The system to be controlled needs to be known only partially: it suffices to predefine the local behaviour around the optimal state and the local shape of the fitness function. For the calculation of optimal responses to time-dependent perturbations, the control and response coefficients were generalised to oscillatory perturbations of finite frequencies. Frequency-dependent control coefficients were defined, and a formula for their calculation as well as summation and connectivity theorems for them were derived. Control coefficients were simulated for a large metabolic network based on a simple assumption about the elasticities. The control coefficients assume almost sparse values and reflect the network topology: distant chemical reactions tend to have a weak influence on each other, and coefficients for a metabolic subsystem

are robust against changes in distant parts of the surrounding system. For the frequency-dependent control coefficients, no strong resonances were found.

The main result of this work is that optimal expression patterns are related to response coefficients. In particular, correlations between protein interaction pairs and within functional modules and functional gene classes were predicted. For metabolic systems, the summation and connectivity theorems of MCA lead to sum rules that relate optimal regulation profiles to the structure of the metabolic network and imply correlated expression. If the assumptions made, namely optimal response and small perturbations, hold for the experiments studied, then the response coefficient profiles should appear as statistical components in the expression data. Between experimental expression data and simulated control coefficients, weak but significant relations were found, thus supporting the hypothesis of optimal regulation. The sparse structure of response coefficients suggests the use of ICA and other statistical methods that are sensitive to sparse components, and provides a function-related interpretation for them. Further predictions concern the compensation for deletions and a relation between differential expression and the diminished growth rate after deletions. If optimal regulation is realised by feedback signals from the cell variables to the regulators, then functional relations should also be portrayed in the linear feedback coefficients, thus functionally related genes are likely to share their input signals.

Explaining expression patterns by their function does not exclude a causal explanation, for instance, by signals from transcription factors binding to sequence motifs. Quite the contrary, from a teleological point of view, the regulation machinery can be regarded as the implementation of an optimal behaviour. Altogether, the optimality-based model complements the causal biochemical approach by claiming a relation between biological function, expression and regulatory mechanism as it is realised, for instance, in operons.

Appendix A

Proofs and additional formulae

A.1 Mathematical symbols

I	identity matrix	matrix
$M, M^{\perp}$	a subspace and its orthogonal space	
$\tilde{X}^{(n)}$	first n principal components of matrix X	matrix
ϵ	elasticities with respect to metabolites	matrix
π_x	elasticities with respect to parameter x	matrix
N	stoichiometric matrix	matrix
K	kernel matrix	matrix
L	link matrix	matrix
N^0	stoichiometric matrix for independent metabolites	matrix
M^0	Jacobian for independent metabolites	matrix
C^S	concentration control coefficients	matrix
C^J	flux control coefficients	matrix
R^S, R^J	response coefficients	matrix
Q	uncorrelated control profiles	matrix

Table A.1: Mathematical symbols for the description of metabolic networks. Vectors are denoted by usual small letters. Capital letters denote matrices or fitness functions. Subscripts that are part of a symbol indicate derivatives.

x	regulatory variables	vector
α	environmental variables	vector
$y(x,\alpha)$	output variables	vector
$(R_x^y)_{ik} = \mathrm{d}y_i/\mathrm{d}x_k$	response coefficients	matrix
$(R_{ab}^y)_{il}^k = \mathrm{d}y_k^2/(\mathrm{d}x_i \mathrm{d}x_l)$	second-order response coefficients with respect to a and b	tensor
$F(x,y)$	fitness function	scalar
$U(\mathrm{x})$, $V(y)$	fitness contribution by x and y	scalar
$F_x = \nabla_x F$	marginal fitness with respect to x	vector
$F_y = \nabla_y F$	marginal fitness with respect to y	vector
$G(x,\alpha) = F(x, y(x,\alpha))$	effective fitness	scalar
$G_x = \nabla_x G$	effective marginal fitness with respect to x	vector
$(G_{xx})_{ik} = \mathrm{d}^2 G/(\mathrm{d}x_i \mathrm{d}x_k)$	effective fitness curvature	matrix
$T_{ab} = F_y^T R_{ab}^y$	eff. fitness curvature by second-order response	matrix
$\mathrm{d}\hat{\alpha}$	perturbation of α	vector
$\mathrm{d}\bar{x}$	optimal response of x	vector
w_y^x	signalling coefficients	matrix
W	regulatory coefficients	matrix

Table A.2: Mathematical symbols used in the studies on optimal expression.

A.2 Derivation of Equation (5.23)

If the output variables y have to change by a fixed amount $\mathrm{d}y = R_x^y \mathrm{d}x$, the condition for optimal $\mathrm{d}\bar{x}$ is

$$F(x + \mathrm{d}\bar{x}, y + \mathrm{d}y) = \max \qquad \text{with the constraint} \quad R_x^y \, \Delta\bar{x} = \Delta y$$

Optimisation using Lagrangian multipliers λ yields

$$\begin{aligned} 0 &= \nabla_{\mathrm{d}\bar{x}} \left[F(x + \mathrm{d}\bar{x}, y + \mathrm{d}y) + \lambda^T (R_x^y \, \mathrm{d}\bar{x} - \mathrm{d}y) \right] \\ &\approx \nabla_{\mathrm{d}\bar{x}} \left[1/2 \, \mathrm{d}\bar{x}^T \, F_{xx} \, \mathrm{d}\bar{x} + \lambda^T (R_x^y \, \mathrm{d}\bar{x} - \mathrm{d}y) \right] \end{aligned} \tag{A.1}$$

where F was expanded to second order and used $F_x + F_y\ R_x^y = 0$, which holds in the initial optimal state. It follows

$$\begin{aligned} 0 &= 1/2\ F_{xx}\ \mathrm{d}\bar{x} + R_x^{y^T}\ \lambda \\ \mathrm{d}\bar{x} &= -2\ F_{xx}^{-1}\ R_x^{y^T}\ \lambda \\ \mathrm{d}y &= R_x^y\ \mathrm{d}\bar{x} = -2\ R_x^y\ F_{xx}^{-1}\ R_x^{y^T}\ \lambda \\ \lambda &= -1/2\ (R_x^y\ F_{xx}^{-1}\ R_x^{y^T})^{-1}\ \mathrm{d}y \\ \mathrm{d}\bar{x} &= F_{xx}^{-1}\ R_x^{y^T} (R_x^y\ F_{xx}^{-1}\ R_x^{y^T})^{-1}\ \mathrm{d}y \end{aligned} \tag{A.2}$$

if F_{xx} is invertible and R_x^y has full row rank.

A.3 Derivation of Equation (5.24)

The optimal regulatory profile $\mathrm{d}\bar{x}$ has to fulfil

$$G_x(x + \mathrm{d}\bar{x}, \alpha) - \lambda\ d\hat{x} = 0 \tag{A.3}$$

where $d\hat{x} = (0\ ..\ 0\ \mathrm{d}\hat{x}_i\ 0\ ..\ 0)^T$ represents the constrained variable i, and λ is a Lagrangian multiplier. We expand

$$G_x(x + \mathrm{d}\bar{x}, \alpha) \approx G_x(x, \alpha) + G_{xx}\ \mathrm{d}\bar{x} \tag{A.4}$$

where $G_{xx} = G_{xx}(x, \alpha)$. As $G_x(x, \alpha) = 0$ for the unperturbed state,

$$d\bar{x} = \lambda\ G_{xx}^{-1}\ d\hat{x} \tag{A.5}$$

From $d\bar{x}_i = \mathrm{d}\hat{x}_i$ follows $\lambda = 1/(G_{xx}^{-1})_{ii}$.

A.4 Derivation of Equation (5.26)

Using the expansion for matrices X with small absolute eigenvalues $|\lambda| < 1$

$$(1 - X)^{-1} = \sum_{n=0}^{\infty} X^n$$

and the identity

$$(A + B)^{-1} = (1 - A^{-1}B)^{-1}A^{-1} = A^{-1}(1 - BA^{-1})^{-1}$$

the expansion is straightforward.

A.5 Derivation of Equation (5.27)

Equation (5.19 yields

$$\mathrm{d}\bar{x} = -(F_{xx} + T_{xx} + R_x^{y^T} F_{yy} R_x^y)^{-1}\ R_x^{y^T}\ \mathrm{d}\hat{F}_y \tag{A.6}$$

It has to be shown that for $F_{xx} + T_{xx} \to 0$, this equals

$$\mathrm{d}\bar{x} = -R_x^{y^+} F_{yy}^{-1}\ \mathrm{d}\hat{F}_y = -R_x^{y^T}(R_x^y\ R_x^{y^T})^{-1}\ F_{yy}^{-1}\ \mathrm{d}\hat{F}_y \tag{A.7}$$

Equating (A.6) and (A.7) yields

$$\begin{aligned} -(F_{xx} + T_{xx} + R_x^{y^T} F_{yy} R_x^y)^{-1}\ R_x^{y^T}\mathrm{d}\hat{F}_y &= -R_x^{y^T}(R_x^y\ R_x^{y^T})^{-1}\ F_{yy}^{-1}\mathrm{d}\hat{F}_y \qquad \text{(A.8)} \\ R_x^{y^T}\mathrm{d}\hat{F}_y &= (F_{xx} + T_{xx} + R_x^{y^T} F_{yy} R_x^y) R_x^{y^T}(F_{yy}\ R_x^y\ R_x^{y^T})^{-1}\mathrm{d}\hat{F}_y \end{aligned}$$

which holds true if $F_{xx} + T_{xx} \to 0$.

A.6 Derivation of Equation (5.33)

Given a perturbation $\mathrm{d}a$, we expand the change in fitness

$$\mathrm{d}^2 G = \begin{pmatrix} G_x \\ G_\alpha \end{pmatrix}^T \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}\alpha \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \mathrm{d}x \\ \mathrm{d}\alpha \end{pmatrix}^T \begin{pmatrix} G_{xx} & G_{x\alpha} \\ G_{\alpha x} & G_{\alpha\alpha} \end{pmatrix} \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}\alpha \end{pmatrix} \tag{A.9}$$

Let us assume normally-distributed perturbations $\mathrm{d}\alpha$ with mean $\langle \mathrm{d}\alpha \rangle = 0$ and covariance matrix $\mathrm{cov}(\mathrm{d}\alpha) = \langle \mathrm{d}\alpha\ \mathrm{d}\alpha^T \rangle$. Considering that the first order-terms vanish on average, and assuming that $\mathrm{d}x$ remains 0, we get

$$\langle \mathrm{d}^2 G \rangle = \frac{1}{2}\ \langle \mathrm{d}\alpha^T\ G_{\alpha\alpha}\ \mathrm{d}\alpha \rangle = \frac{1}{2}\ Tr(G_{\alpha\alpha}\ \mathrm{cov}(\mathrm{d}\alpha)) \tag{A.10}$$

On the other hand, inserting the optimal response $\mathrm{d}\bar{x} = -G_{xx}^{-1}\ G_{x\alpha}\ \mathrm{d}\alpha$ yields

$$\langle \mathrm{d}^2 \bar{G} \rangle = \frac{1}{2}\ \langle \mathrm{d}\alpha^T (G_{\alpha\alpha} - G_{\alpha x}\ G_{xx}^{-1}\ G_{x\alpha}) \mathrm{d}\alpha \rangle \tag{A.11}$$

Due to regulation, the expected fitness increases by

$$\langle \bar{G} - \hat{G} \rangle = \langle \mathrm{d}^2 \bar{G} - \mathrm{d}^2 \hat{G} \rangle = -\frac{1}{2}\ Tr(\ G_{\alpha x}\ G_{xx}^{-1}\ G_{x\alpha}\ \mathrm{cov}(\mathrm{d}\alpha)) \tag{A.12}$$

A.7 Derivation of Equation (5.30)

The fitness $G(x, \alpha)$ is maximal before and after the perturbation, therefore $\mathrm{d}\bar{G}_x$ has to vanish. From

$$0 = \mathrm{d}\bar{G}_x \quad = \quad \mathrm{d}\bar{F}_x + R_x^{y^T}\mathrm{d}\bar{F}_y \tag{A.13}$$

and $\mathrm{d}\bar{F}_x = F_{xx}\ \mathrm{d}\bar{x}$ follows

$$\mathrm{d}\bar{x} \quad = \quad -F_{xx}^{-1}R_x^{y^T}\mathrm{d}\bar{F}_y \tag{A.14}$$

Note that in equation A.13 we neglected the term $\mathrm{d}\bar{R}_x^{y^T}F_y$.

A.8 Derivation of the projector property (6.7)

P^y reads

$$P^y = R_x^y G_{xx}^{-1} R_x^{y^T} F_{yy} = R_x^y (F_{xx} + R_x^{y^T} F_{yy} R_x^y)^{-1} R_x^{y^T} F_{yy} \tag{A.15}$$

This matrix maps any vector to $\mathrm{span}(R_x^y)$, while for $F_{xx} \to 0$, R_x^y is mapped to itself. Accordingly, the transposed maps any vector to $\mathrm{span}(F_{yy}R_x^y)$, and $F_{yy}R_x^y$ is mapped to itself. It is also easy to see that P^y fulfils the projector property $P = P^2$.

A.9 Effective fitness for constrained regulation

The derivatives of the effective fitness H (see section 6.2) reflect the derivatives of F and A. With the tensor notation $(H_a)_i = \partial_{a_i} h$ and $(H_{ab})_{ik} = \partial_{a_i}\partial_{b_k} H$ (similar for derivatives

of F), and $(R^z_a)^l_i = \partial_{a_i} A^l$, $(R^z_{ab})^l_{ik} = \partial_{a_i}\partial_{b_k} A^l$, they read

$$
\begin{aligned}
H_x &= F_x \\
(H_y)_i &= (F_z)_k \; (R^z_y)^k_i \\
(H_\alpha)_i &= (F_z)_k \; (R^z_\alpha)^k_i \\
H_\beta &= F_\beta \\
H_{xx} &= F_{xx} \\
(H_{xy})_{ik} &= (F_{xz})_{il} \; (R^z_y)^l_k \\
(H_{x\alpha})_{ik} &= (F_{xz})_{il} \; (R^z_\alpha)^l_k \\
H_{x\beta} &= F_{x\beta} \\
(H_{yy})_{ik} &= (R^z_y)^l_i \; (F_{zz})_{lm} \; (R^z_y)^m_k + (F_z)_l \; (R^z_{yy})^l_{ik} \\
(H_{y\alpha})_{ik} &= (R^z_y)^l_i \; (F_{zz})_{lm} \; (R^z_\alpha)^m_k + (F_z)_l \; (R^z_{y\alpha})^l_{ik} \\
(H_{y\beta})_{ik} &= (R^z_y)^l_i \; (F_{z\beta})_{lk} \\
(H_{\alpha\alpha})_{ik} &= (R^z_\alpha)^l_i \; (F_{zz})_{lm} \; (R^z_\alpha)^m_k \\
(H_{\alpha\beta})_{ik} &= (R^z_\alpha)^l_i \; (F_{z\beta})_{lk} \\
H_{\beta\beta} &= F_{\beta\beta}
\end{aligned}
$$

Appendix B

Expression data and analysis methods

B.1 Parameters for artificial expression data

Parameter	Distribution	(1)	(2)	(3)	(4)
samples			100	200	50
genes			1	50	500
factors			3	5	5
regulated			all	all	300
input			all	all	3
$p(w)$	normal	0 ± 0.4	0 ±0.5	0 ± 0.4	0 ± 0.4
$p(w_0)$	normal	0 ± 0.5	0 ±0	0 ± 0.5	0 ± 0.5
$p(x_{max})$	log normal	0 ± 1	0 ±0	0 ± 1	0 ± 1
$p(\eta_{mult})$	log normal	0 ± 0.15	0 ±0.15	0 ± 0.15	0 ± 0.15
$p(\eta_{add})$	log normal	-1.5 ± 0.2	-2 ±0.2	-1.5 ± 0.2	-1.5 ± 0.2

Table B.1: Parameter values for simulated data. The parameters sets were used in the calculations for (1) Figure 2.3, (2) Figure 2.5, (3) Figure 3.1, (4) Figure 4.2. Distributions are characterised by their means and standard deviations of the variable itself (for normal distributions) or its natural logarithm (for log-normal distributions).

	weakly nonlinear	nonlinear
standard		$p(w) = 0 \pm 0.2$
no noise	$p(\eta_{add}) = -10 \pm 0$ $p(\eta_{mult}) = 0 \pm 0$	$p(\eta_{add}) = -10 \pm 0$ $p(\eta_{mult}) = 0 \pm 0$ $p(w) = 0 \pm 0.2$
strong noise	$p(\eta_{mult}) = 0 \pm 0.2$ $p(\eta_{add}) = -2 \pm 0.2$	$p(\eta_{mult}) = 0 \pm 0.2$ $p(\eta_{add}) = -2 \pm 0.2$ $p(\eta_{mult}) = 0 \pm 0.2$
lower saturation	$p(w_0) = -1 \pm 0$	$p(w_0) = -1 \pm 0$ $p(w) = 0 \pm 0.2$
upper saturation	$p(w_0) = 1 \pm 0$	$p(w_0) = 1 \pm 0$ $p(w) = 0 \pm 0.2$

Table B.2: Variation of parameter values for simulated data. The parameter sets correspond to the boxes of Figure 4.4 and to the Figures 2.3 and C.2

B.2 Data used in this work

The publicly available sets of microarray data used in this thesis, are listed in Table B.3). In the experiments, cDNA (reverse-transcribed mRNA) populations from the sample being studied and from a reference sample were stained with different fluorescent dyes and both hybridised to the same chip. The gene expression matrix X contains the log-ratios $X_{ik} = \log_2(R_{ik}/G_{ik})$ between the red (experiment) and green (reference) intensities. As the mean values for genes and samples depend strongly on the hybridisation procedure and on data normalisation, they were shifted to zero. The missing values were replaced by zeros afterwards. Gene profiles may be visualised as a cloud of points in a n-dimensional space (where n is the number of samples). Centring within genes shifts to centre of mass of the cloud to the origin, while centring within samples projects the cloud to a hyperplane orthogonal to the vector $(1, 1, .., 1, 1)^T$.

B.3 Algorithms

Most of the calculations in this work were done in MATLAB. I used several publicly available MATLAB packages, listed in Table B.4.

The independent components were standardised according to the following conventions:

	Experiment	Samples	Reference
Yeast	Cell cycle	77	Spellman et al. [107] cellcycle-www.stanford.edu
Yeast	Stress response	45	Causton et al. [10] web.wi.mit.edu/young/environment/
Yeast	Stress response	174	Gasch et al. [32] genome-www.stanford.edu/yeast_stress/data.shtml
Yeast	Deletions and drugs	300	Hughes et al. [51] www.rii.com/tech/pubs/cell_hughes.htm
Yeast	Deletions in galactose pathway	24	Ideker et al. [61] www.sciencemag.org/cgi/content/full/292/5518/929/DC1
Human	Lymphocytes, lymphoma	95	Alizadeh et al. [2] llmpp.nih.gov/lymphoma/
Yeast	Growth of deletion mutants		Giaever et al. [35] genomics.lbl.gov/YeastFitnessData/websitefiles/cel_index.html
Yeast	Binding of transcription factors	113	Lee et al. [76] staffa.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=17&f=downloaddata

Table B.3: Experimental data sets studied

Method	Publication	Source for MATLAB code
FastICA	[59]	www.cis.hut.fi/projects/ica/fastica/
NNMF	[75]	
Topographic ICA	[56]	www.cis.hut.fi/projects/ica/imageica/
Nonlinear ICA	[72]	www.cis.hut.fi/projects/ica/bayes/
Canonical analysis	[100]	
K means clustering		www-2.cs.cmu.edu/~dellaert/software/
Multi-layer-perceptron	[88]	www.ncrg.aston.ac.uk/netlab/

Table B.4: Sources of some algorithms used in this work

1. When compared to components representing noise, biological components should be more informative, showing a large contrast J_G, and they should also capture a higher amount J_A of the data variance. With centred data and components scaled to unit variance, the variance explained by component k is proportional to $J_A(k) = \sum_l A_{kl}^2$. To take into account both properties, and without considering a biological meaning behind the exact order, the components were sorted according to a linear combination

$$s^{(k)} = c\ J_G^{(k)} / < J_G > + (1-c)\ J_A^{(k)} / < J_A >$$

of both quantities, scaled by their mean values, with some arbitrary $c \in [0, 1]$.

2. For each component, the sign was chosen such that the mean influence was higher than the median. Accordingly, a mode will rather induce than repress genes, which is of course not more than a convention: when a mode is downregulated, the genes repressed by it are upregulated.

3. By setting the gene mean values to zero, the mean values of the modes were implicitly shifted to zero as well.

4. For comparing different simulation runs, order and signs of the components were chosen by matching the modes to the modes of a reference model. To do so the sum of squared correlations was maximised by simulated annealing.

B.4 Mapping yeast ORF to the metabolic network

The metabolic network was build from the LIGAND database [64]. To plot expression data on the Boehringer map [83] and to compare them to response coefficients, yeast ORF were mapped to chemical reactions. Both ORF and reactions were mapped to EC numbers: virtual EC expression data and EC response coefficients were defined by averaging over ORF and reactions related to the same EC number, respectively. A network distance between EC numbers was defined by the smallest distance between corresponding chemical reactions. To visualise values related to chemical reactions, a dot is drawn on the Boehringer map for any reaction with this EC number. Therefore, dots may appear for reactions that are not present in yeast, and plotting an enzyme for one reaction results in multiple dots. The Boehringer map was used with the permission of Spektrum Akademischer Verlag, GmbH.

Appendix C

Additional tables and figures

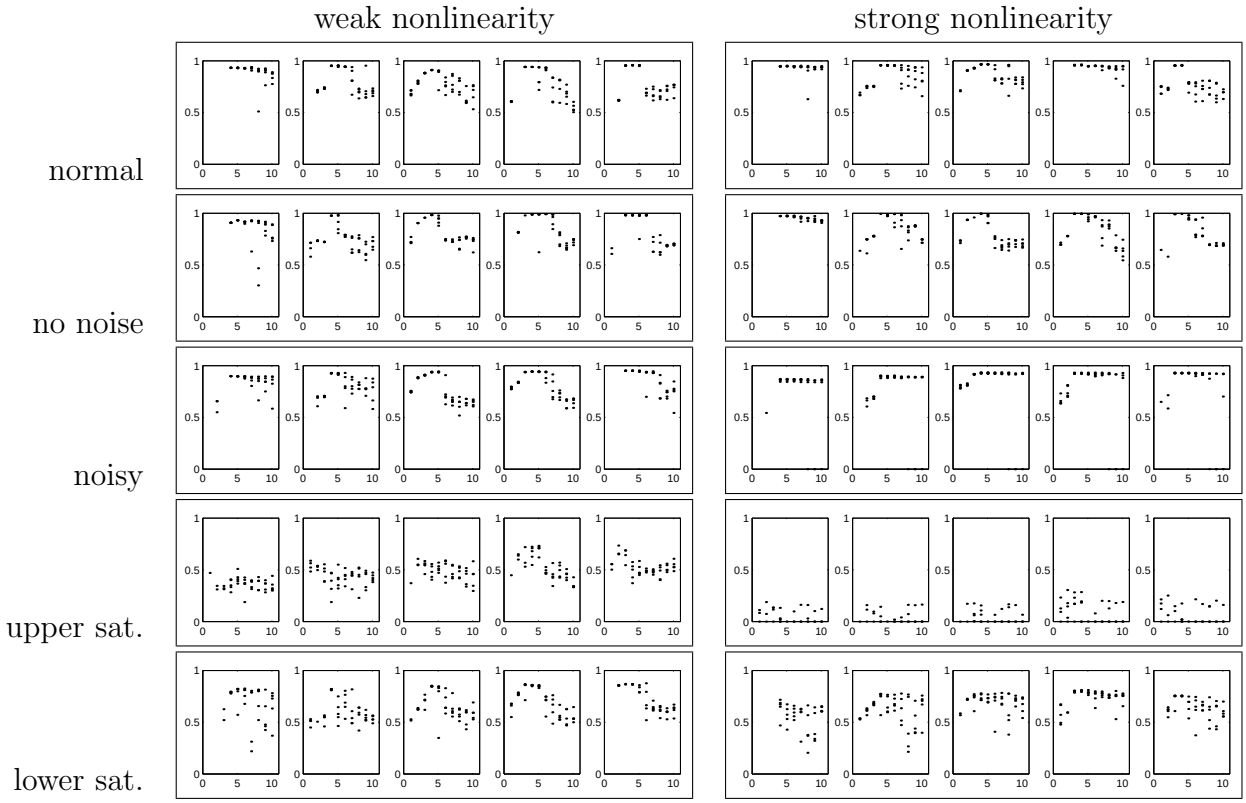


Figure C.1: Reconstruction of expression modes, for different parameter settings (see Table B.2). The diagrams are analogous to those in Figure 4.4. Here, correlations of input weights are shown.

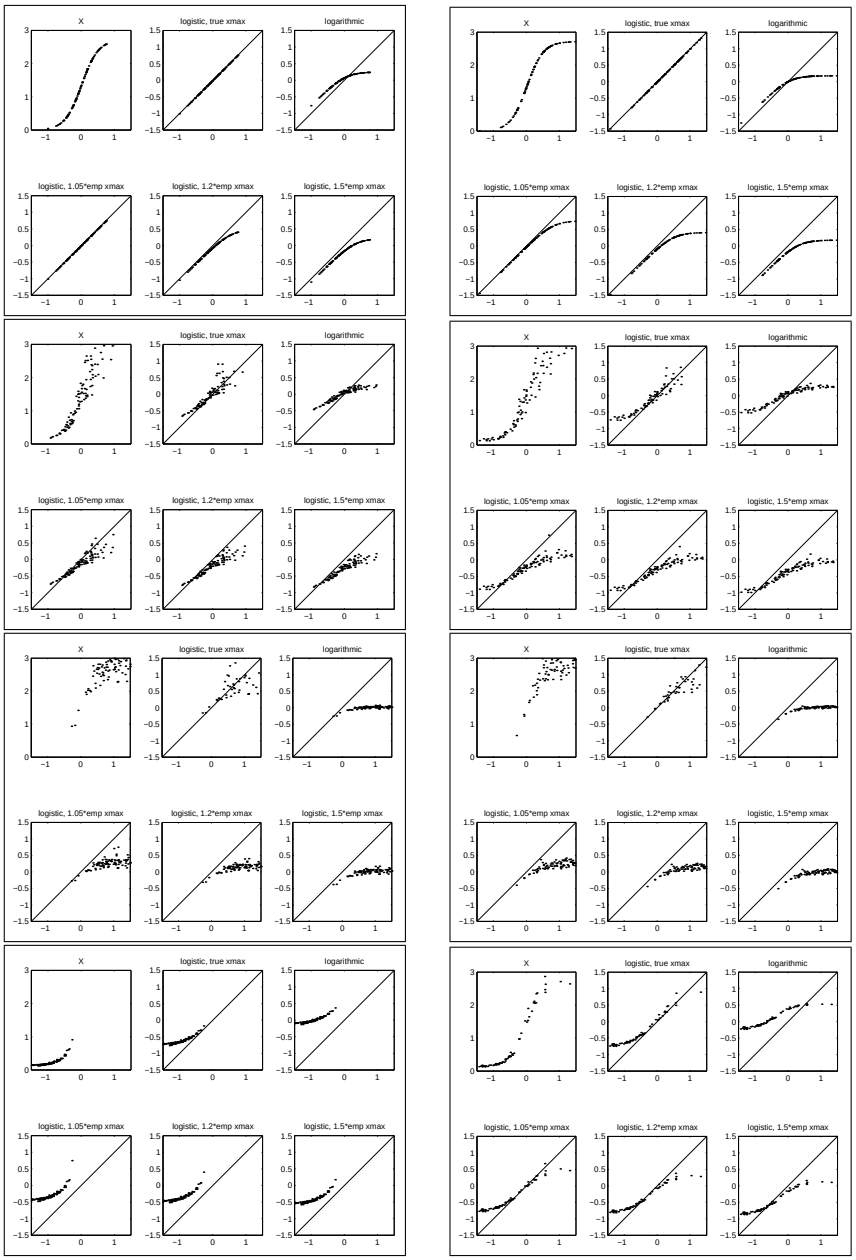


Figure C.2: Transformations of artificial data with different parameter sets. Each row of large boxes is analogous to Figure 2.3. The four rows show the effects of vanishing noise, stronger noise, and data points in the upper or lower saturation region. For most conditions, the logistic transformation performs equally well or slightly better than the log-transformation

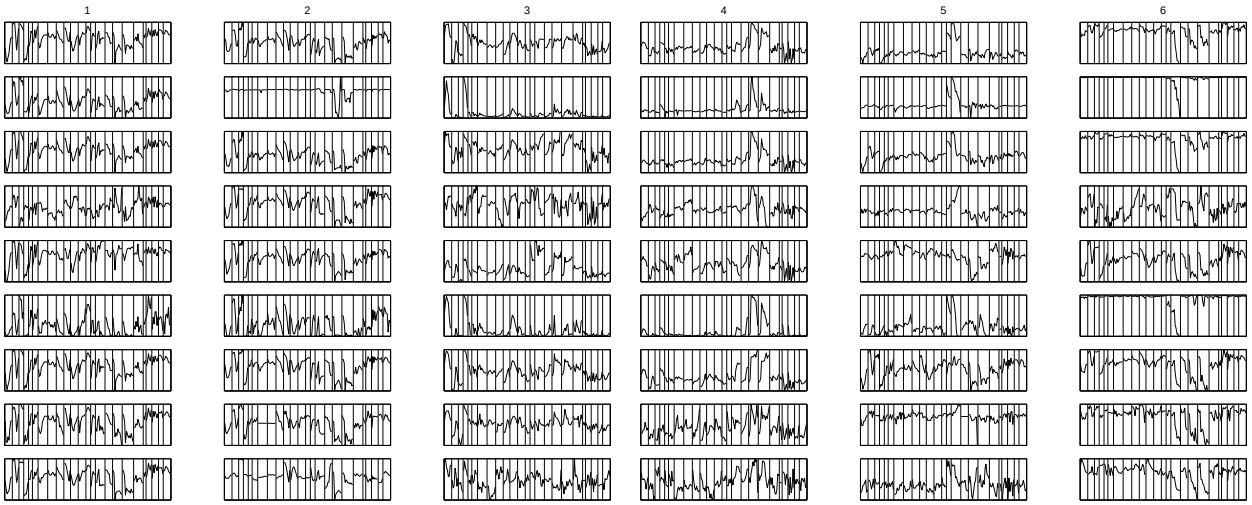


Figure C.3: Linear models for the stress response data Gasch et al. [32]. The models (shown in the different rows) are the same as in Figure 4.9. The remaining 6 modes are shown in Figure C.4. The experimental time courses show reactions to different kinds of experimental perturbations, some of which are listed in Figure 7.1.

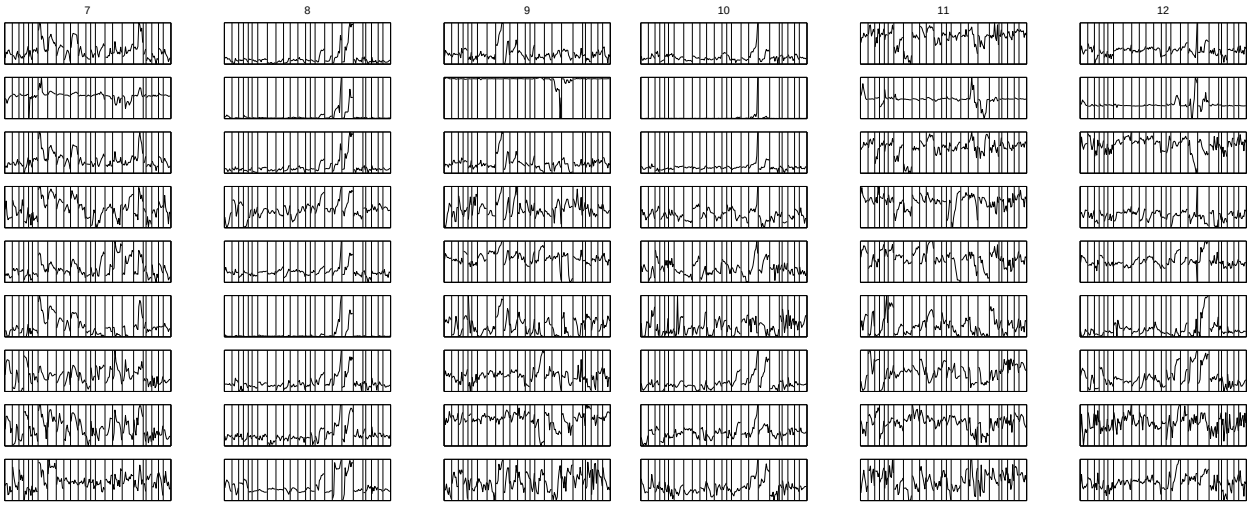


Figure C.4: Linear models for the stress response data Gasch et al. [32]. Same as Figure C.3: the remaining six components are shown.

	Mode	m	Genes	Functional category
ICA	2	16	155	CYTOPLASM
	4	11	59	MITOCHONDRION
	7	7.4	18	DETOXIFICATION
	9	7.3	14	ENDOPLASMIC RETICULUM
	1	6.1	10	RRNA TRANSCRIPTION
	1	5.6	175	UNCLASSIFIED PROTEINS
	7	5.5	8	FERMENTATION
	2	5.4	8	RIBOSOME BIOGENESIS
	5	5.4	16	AMINO ACID METABOLISM
	1	5.3	57	NUCLEUS
	6	4.5	12	NUCLEOTIDE METABOLISM
	5	3.4	7	NITROGEN AND SULFUR METABOLISM
	5	3.1	23	PLASMA MEMBRANE
PCA	7	6.6	8	FERMENTATION
	7	5.3	12	DETOXIFICATION
	4	4.8	22	MITOCHONDRION
	9	3.7	6	ENDOPLASMIC RETICULUM
	5	3.4	12	PLASMA MEMBRANE
	11	3	32	CYTOPLASM
TICA	2	12	134	CYTOPLASM
	4	7.6	44	MITOCHONDRION
	5	6.9	41	MITOCHONDRION
	1	6.4	10	RRNA TRANSCRIPTION
	7	6	13	DETOXIFICATION
	2	5	9	RIBOSOME BIOGENESIS
	7	5	7	FERMENTATION
	1	4	102	UNCLASSIFIED PROTEINS
	3	3.5	14	AMINO ACID METABOLISM
NNMF	2	7.9	105	CYTOPLASM
	5	6.6	13	AMINO ACID METABOLISM
	4	5.1	20	MITOCHONDRION
	6	5	14	PLASMA MEMBRANE
	1	4.1	16	NUCLEUS
	2	3.6	8	RRNA TRANSCRIPTION
	2	3.4	7	RIBOSOME BIOGENESIS
	7	3	6	DETOXIFICATION
K means	2	20	148	CYTOPLASM
	1	7.6	11	RRNA TRANSCRIPTION
	4	6.9	10	PEROXISOME
	10	5	31	ENDOPLASMIC RETICULUM
	2	4.8	6	RIBOSOME BIOGENESIS
	10	4.7	16	INTRACELLULAR TRANSPORT VESICLES
	10	4.7	19	GOLGI
	1	4.6	121	UNCLASSIFIED PROTEINS
	12	4.5	234	UNCLASSIFIED PROTEINS
	10	3.9	21	CYTOSKELETON
	1	3.4	42	NUCLEUS
	12	3.3	9	OTHER TRANSPORT FACILITATORS
	10	3.3	12	PROTEOLYTIC DEGRADATION
	6	3.2	23	ENDOPLASMIC RETICULUM
Canonical	9	6.6	23	MITOCHONDRION
	12	4.3	7	AMINO ACID METABOLISM
	12	3.9	7	NUCLEOTIDE METABOLISM
	5	3.8	11	PLASMA MEMBRANE
canonical	4	3.4	11	MITOCHONDRION

Table C.1: Correspondence between expression modes from cell stress data Gasch et al. [32] and functional categories. Compare Table 4.2 .

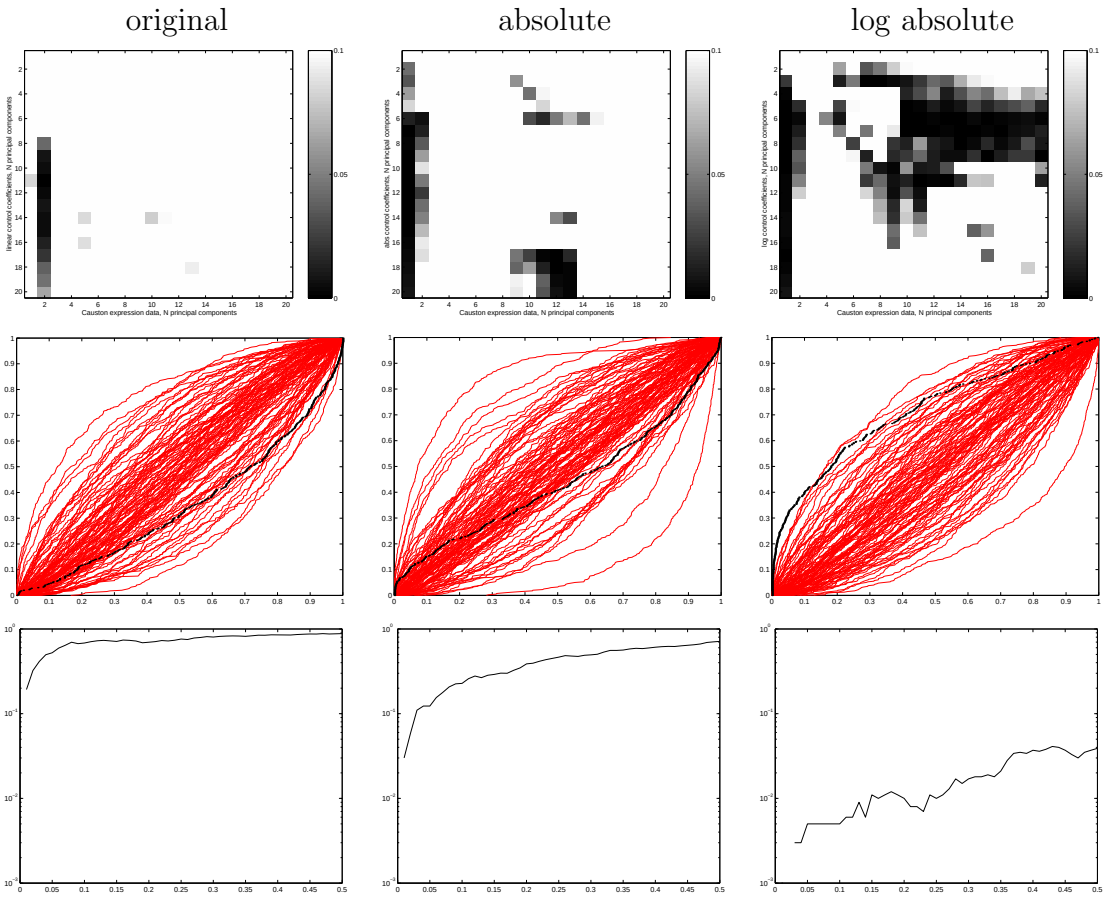


Figure C.5: Linear relation between response coefficients and expression data. Same as Figure 9.3, for stress response data Causton et al. [10].

Bibliography

[1] B. Alberts et al. *Molecular biology of the cell*. Garland Publishing, Inc., 3. edition, 1994.

[2] A.A. Alizadeh, M.B. Eisen, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] C. Allen. Teleological notions in biology. In *The Stanford Encyclopedia of Philosophy*. 1999. http://plato.stanford.edu/archives/sum1999/entries/teleology-biology/.

[4] O. Alter, P.O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 97(18):10101–10106, 2000.

[5] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6:281–297, 1999.

[6] A. Brazma et al. Predicting gene regulatory elements in silico on a genomic scale. *Genome research*, 8:1202, 1998.

[7] M.P.S. Brown et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.*, 97:262–267, 2000.

[8] F. Bruggeman et al. Modular response analysis of cellular regulatory networks. *J. theor. Biol.*, 218:507–520, 2002.

[9] H.J. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

[10] H.C. Causton et al. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the cell*, 12:323–337, 2001.

[11] K.C. Chen et al. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Molecular Biology of the Cell*, 11:369–391, 2000.

[12] R. Cho et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[13] J.M. Claverie. Computational methods for the identification of differential and coordinated gene expression. *Human molecular genetics*, 8:1821, 1999.

[14] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[15] A. Cornish-Bowden and M.L. Cárdenas. Channeling can affect concentrations of metabolic intermediates at constant net flux: artefact or reality? *Eur. J. Biochem.*, 213:87–92, 1993.

[16] T. M. Cover and J. A. Thomas. *Elements of information*. Wiley, 1991.

[17] A. de la Fuente et al. Linking the genes: inferring quantitative gene networks from microarray data. *Trends in genetics*, 18(8):395–398, 2002.

[18] J.L. DeRisi et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[19] P. D'Haeseleer et al. Mining the gene expression matrix: Inferring gene relationships from large-scale gene expression data. *Information Processing in Cells and Tissues, Holcombe and Paton Eds., Plenum Press, N.Y.*, 1998.

[20] P. D'Haeseleer et al. Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing '99, World Scientific Publishing*, 1999.

[21] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. 1999.

[22] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Wiley, 2 edition, 2001.

[23] S. Dudoit et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, August 2000.

[24] J.S. Edwards, R. Ramakrishna, and B.O. Palsson. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng.*, 77(1):27–36, 2002.

[25] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[26] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, 1998.

[27] D. Fambrough et al. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather then independent, sets of genes. *Cell*, 97:727, 1997.

[28] K. Fellenberg et al. Correspondence analysis applied to microarray data. *PNAS*, 98(19):10781–10786, 2001.

[29] A. Fire. RNA-triggered gene silencing. *Trends Genet.*, 15(9):358–963, 1999.

[30] J. Förster et al. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Research*, 13(2):244–253, 2003.

[31] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. In *Proceedings of the fourth annual international conference on computational molecular biology on RECOMB 2000*, pages 127–135, 2000.

[32] A.P. Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

[33] A. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[34] A. Gelman, J. B. Carlin, H. S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman & Hall, 1997.

[35] G. Giaever et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418:387–391, 2002.

[36] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Research*, 29(17):3513–3519, 2001.

[37] S.P. Gygi et al. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730, 1999.

[38] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(90001):145S–154, 2002.

[39] R. Heinrich. Mathematical models of metabolic systems: general principles and control of glycolysis and membrane transport in erythrocytes. *Biomed. Biochim. Acta*, 44(6):913–927, 1985.

[40] R. Heinrich, H.G. Holzhütter, and S. Schuster. A theoretical approach to the evolution and structural design of enzymatic networks; Linear enzymatic chains, branched pathways and glycolysis of erythrocytes. *Bull. Math. Biol.*, 49:539–595, 1987.

[41] R. Heinrich and S. Schuster. *The regulation of cellular systems.* Chapman & Hall, 1996.

[42] R. Heinrich and S. Schuster. The modelling of metabolic systems. structure, control, and optimality. *BioSystems*, 47:61–77, 1998.

[43] J.v. Helden et al. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.

[44] J. Henderson and R. Quandt. *Microeconomic Theory: A Mathematical Approach.* New York: McGraw-Hill, 3rd ed. edition, 1980.

[45] T. Höfer and R. Heinrich. A second-order approach to metabolic control analysis. *J. theor. Biol*, 1993.

[46] F. Holstege et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717–728, 1998.

[47] N.S. Holter et al. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *PNAS*, 97(15):8409–8414, 2000.

[48] N.S. Holter et al. Dynamic modelling of gene expression data. *PNAS*, 98(4):1693–1698, 2001.

[49] G. Hori et al. Blind gene classification - an application of a signal separation method. *Genome Informatics*, 12:255–256, 2001.

[50] W. Huber et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 1(1):1–9, 2002.

[51] T.R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[52] F. Hynne, S. Danø, and P.G. Sørensen. Full-scale model of glycolysis in Saccharomyces cerevisiae. *Biophysical Chemistry*, 94:121–163, 2001.

[53] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[54] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, pages 94–128, 1999.

[55] A. Hyvärinen and P.O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[56] A. Hyvärinen, P.O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1525–1558, 2001.

[57] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[58] A. Hyvärinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, in press.

[59] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[60] R.U. Ibarra, J.S. Edwards, and B.O. Palsson. Escherichia Coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420:186–189, 2002.

[61] T. Ideker et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292:929–934, 2001.

[62] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12(1):37–46, 2002.

[63] D. Kahn and H.V. Westerhoff. Control theory of regulatory cascades. *J. theor. Biol.*, 153:255–285, 1991.

[64] M. Kanehisa, S. Goto, et al. The KEGG databases at genomenet. *Nucleic Acids Res.*, 30:42–46, 2002.

[65] J. Karhunen. *ICA: Principles and practice*, chapter Nonlinear independent component analysis. 2000.

[66] S. A. Kauffman. *What is life? The next fifty years*, chapter What is life?: Was Schrödinger right? Cambridge University press, 1995.

[67] E. Klipp and R. Heinrich. Competition for enzymes in metabolic pathways: implications for optimal distributions of enzyme concentrations and for the distribution of flux control. *BioSystems*, 54:1–14, 1999.

[68] E. Klipp, R. Heinrich, and H.G. Holzhütter. Prediction of temporal gene expression. Metabolic optimization by re-distribution of enzyme activities. *Eur. J. Biochem*, 269:1–8, 2002.

[69] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[70] J.R. Koza. *Genetic Programming*. Complex adaptive systems. MIT press, 1998.

[71] M.P. Kurhekar et al. Genome-wide pathway analysis and visualization using gene expression data. In *Pacific Symposium on Biocomputing*, volume 7, pages 462–473, 2002.

[72] H. Lappalainen. Nonlinear independent component analysis using ensemble learning: theory. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2000*, pages 251–256, 2000.

[73] H. Lappalainen et al. Nonlinear independent component analysis using ensemble learning: experiments and discussion. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2000*, pages 351–356, 2000.

[74] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2002.

[75] D.D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788, 1999.

[76] T.I. Lee et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[77] J.W. Lengeler, G. Drews, and H.G. Schlegel, editors. *Biology of the Prokaryotes*. Thieme Stuttgart, 1999.

[78] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60, 2002.

[79] W. Liebermeister et al. A theory of optimal differential gene expression. *to appear in BioSystems*, 2004.

[80] E.M. Marcotte et al. A combined algorithm for genome-wide prediction and function. *Nature*, 402:83–86, 1999.

[81] F. Mensonides et al. The metabolic response of Saccharomyces cerevisiae to continuous heat stress. *Mol. Biol.Reports*, 29:103–106, 2002.

[82] H. W. Mewes et al. MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Research*, 27:44,48, 1999.

[83] G. Michal, editor. *Biochemical pathways*. Spektrum Akademischer Verlag, 1999.

[84] J.W. Miskin. *Ensemble learing for independent component analysis*. PhD thesis, Selwyn College, Cambridge, 2000.

[85] T.D. Moloshok et al. Application of Bayesian decomposition for analysing microarray data. *Bioinformatics*, 18(4):566–575, 2002.

[86] E. Morett et al. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature Biotechnology*, 2003. in press.

[87] T.J. Sejnowski M.S. Lewicki. Learning overcomplete representations. unpublished.

[88] I. Nabney. *Netlab: Algorithms for Pattern Recognition*. Advances in Pattern Recognition. Springer.

[89] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann publishers, Inc., 1988.

[90] D. Pe'er et al. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 1:1–9, 2001.

[91] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *PNAS*, 96(8):4285–4288, 1999.

[92] Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, 29(2):153–159, 2001.

[93] L.S. Pontryagin et al. *The mathematical theory of optimal processes*. John Wiley, New York, 1962.

[94] C. Reder. Metabolic control: a structural approach. *J. theor. Biol.*, 135:175–201, 1988.

[95] J.G. Reich. Zur Ökonomie im Proteinhaushalt der lebenden Zelle. *Biomed. Biochim. Acta*, 42(7/8):839–848, 1983.

[96] K. Reijenga et al. Control analysis for autonomously oscillating biochemical networks. *Biophys. Journal*, 82:99–108, 2002.

[97] M. Rizzi et al. In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae: II. Mathematical model. *Biotechnology and Bioengineering*, 1997.

[98] Y. Rodriguez et al. Equivalence of branched and unbranched michaelian pathways concerning periodic signal transmission. *Mol. biol. reports*, 29:63–66, 2002.

[99] J. Rung et al. Building and analysing genome-wide gene disruption networks. *Bioinformatics*, 18(Suppl. 2):202–210, 2002.

[100] G. Saporta. *Probabilités, analyse des données et statistique*. Éditions Technip, 1990.

[101] J.M. Savinell and B.O. Palsson. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J. theor. Biol.*, 154:421–454, 1992.

[102] J. Schuchhardt et al. Normalization strategies for cDNA arrays. *Nucleic Acids Research*, 28(10):E47, 2000.

[103] S. Schuster, D. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.*, 18:326–332, 2000.

[104] S. Schuster, S. Klamt, et al. Use of network analysis of metabolic systems in bioengineering. *Bioproc. Biosyst. Eng.*, 24:363–372, 2002.

[105] D. Segrè, D. Vitkup, and G.M. Church. Analysis of optimality in natural and perturbed metabolic networks. *PNAS*, 99(23):15112–15117, 2002.

[106] K. Simek and M. Kimmel. A note on estimation of dynamics of multiple gene expression based on singular value decomposition. *Math. Biosciences*, 182:183–199, 2003.

[107] P.T. Spellman, G. Sherlock, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[108] L. Stryer. *Biochemistry*. Freeman, New York, 4th edn. edition, 1995.

[109] P. Tamayo, D. Slonim, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, 96:2907–2912, 1999.

[110] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, Suppl. 1:S136–S144, 2002.

[111] S. Tavazoie et al. Systematic determination of genetic network architecture. *Nature genetics*, 22:281–285, 1999.

[112] P. Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[113] M. Wahde and J. Hertz. Coarse-grained reverse engineering of genetic regulatory networks. *BioSystems*, 55:129–136, 2000.

[114] D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. *Pacific symosium on biocomputing*, 4:112–123, 1999.

[115] I. Yanai, A. Derti, and C. DeLisi. Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. *PNAS*, 98(14):7940–7945, 2001.

[116] C.H. Yuh, H. Bolouri, and E.H. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, 1998.

[117] X. Zhou, M.C.J. Kao, and W.H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *PNAS*, 9(20):12783–12788, 2002.

[118] A. Zien, R. Küffner, et al. Analysis of gene expression data with pathway scores. In R. Altman et al., editors, *ISMB00*, pages 407–417, La Jolla, CA, August 2000. AAAI.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig ohne fremde Hilfe verfaßt und nur die angegebene Literatur und Hilfsmittel verwendet zu haben.

Wolfram Liebermeister
21. Mai 2003